A U T O M A T I C A C O U S T I C A N A L Y S I S

O F W A V E F O R M P E R T U R B A T I O N S

STEVEN MARK HILLER

Thesis submitted for the degree of Ph.D.

University of Edinburgh

1985

This thesis is dedicated to my mother

A C K N O W L E D G M E N T S

In the course of preparing this thesis, I have received a considerable amount of support from many individuals and I would like to take this opportunity to offer them my sincerest thanks.

The seed of this thesis was planted by Professor John Laver, of the Department of Linguistics, University of Edinburgh. John is a fine scholar and superb friend who nutured and encouraged me throughout the years of this investigation into the acoustic nature of the pathologic voice. I look forward to our continued collaboration in the study of speech.

Many thanks to my colleague Janet Mackenzie whose knowledge and experience in the speech pathology area helped to refine the techniques developed in this thesis. Janet is a true friend and an excellent sounding board on matters academic and otherwise. I am also grateful for the scientific assistance and friendship of Dr. Robert Hanson, of the Digital Sound Corporation in Santa Barbara, California, who acted as a Visiting Senior Scientist on the Medical Research Council project, and of Eddie Rooney, my other colleague on the MRC project.

A number of people in the Department of Linguistics provided valuable technical support and advice. Norman Dryden and Irene Macleod helped to remove various and assorted bugs from the software while the technical assistance of Stewart Smith and Jeff Dodds kept the electrons flowing in the hardware. Statistical issues were clarified through the kind advice of Dr. Ellen Bard and Dr. Mark Terry.

Large helpings of moral support were dished out by my good friends Sandy Hutcheson, Douglas Mac Lean, Earl Collison, Al Beck and Tara Casey.

Much of the voice data analyzed in Chapter 6 was acquired through the assistance of two outpatient voice clinics, namely the Radcliffe Infirmary in Oxford, supervised through the collaboration of Dr. T. Harris and S. Collins, LCST, and of Dr. A. Maran of the Department of Otolaryngology, Royal Infirmary, Edinburgh.

I am indebted to all these people. However, I take complete responsibility for the final version of this thesis.

Several published articles have appeared in relation to this thesis and these papers can be found in the appendices at the end of this text.

D E C L A R A T I O N

This thesis was composed by myself, being an original and substantial contribution to the work of a research group investigating acoustic aspects of the voice (Medical Research Council Project Grant No. 98207136N 'Acoustic Analysis of Voice Features'). I was responsible for the development and evaluation of all the computer programs which form the acoustic analysis system, in my capacity as a Research Associate employed on the above project.

Steven Hiller

September 1985

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

(The figures and tables listed here appear after the indicated page numbers.)

A B S T R A C T

Over the last two decades, researchers in various fields such as laryngology, speech pathology, speech science and phonetics have demonstrated a growing interest in the acoustic characterization of healthy and pathological voices. This research activity has been the response to the need for objective quantitative techniques for the assessment of laryngeal function. The emphasis of this thesis is the evaluation of laryngeal behavior via the acoustic analysis of irregularities in the periodic structures of speech signals. The purpose of this study was the development of a computer-based system which can differentiate between groups of healthy and pathological speakers by the acoustic analysis of speech waveform perturbations of fundamental frequency and amplitude evidenced in samples of connected speech.

The perturbation measurement system consists of three major components including:

1) Pitch Detection Algorithm -- a modified parallel processor operating in the time domain extracts fundamental frequency and amplitude contours from samples of connected speech. A typological review of pitch detection algorithms presented in Chapter 2 supports the use of a multichannel solution based on temporal structural analysis for processing the wide variety of signals produced during connected speech, especially by speakers with pathological conditions of the larynx. Chapter 3 presents the details of implementing the parallel processor. Particular emphasis is given to the production of accurate pitch extraction results.

2) Non-linear Smoothing Algorithm — a smoother comprised of a running-median and Hanning window is applied to the contours to

produce trend lines of fundamental frequency and amplitude, from which excursions of the unsmoothed values may be extracted. In Chapter 4, a review of the perturbation literature reveals that most studies of laryngeal pathology examined cycle-to-cycle perturbations in samples of sustained phonations while very few investigations have evaluated perturbations on a trend line basis in samples of connected speech. The implementation of the non-linear smoother is discussed in detail in Chapter 5.

3) Statistical Evaluation of Long-term Parameters of Intonation and Perturbation -- the excursions of fundamental frequency and amplitude are statistically evaluated to produce the perturbation parameters. Other special parameters of waveform perturbation are also calculated at this stage. Long-term parameters of intonation are derived from the trend line of fundamental frequency values.

A series of experiments are presented in Chapter 6 which evaluates the performance of the perturbation measurement system. The general conclusion drawn from these experiments is that the intonation and perturbation parameters as extracted by the system are useful for differentiating between groups of healthy speakers and speakers with known pathological conditions of the larynx.

Future research should examine three potential applications of the system for quantifying laryngeal function, including screening, differential diagnosis and tracking changes in the status of laryngeal pathology.

The listings of the relevant programs are not included in the material of this dissertation. They will however be made available for inspection during the oral examination.

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

1.0 INTRODUCTION

The information to be presented in this thesis is set within a speech science context. Speech science is a discipline that necessarily interfaces with a number of other subjects -- speech science contributes as well as receives valuable information from its associated areas. Phonetics acts as the interface between the fields of linguistics and speech science. Phoneticians use speech science techniques to derive quantified phonetic data and in collaboration with speech scientists develop and refine these techniques, making them more phonetically relevant. Similarly, the interface between the engineering and speech sciences is based on the development and application of signal processing techniques which are appropriate to speech signals. The engineering sciences provide expertise in the creation of accurate and efficient signal processing methods and, in return, speech science offers practical areas of application for these techniques. The techniques acquired through the interfaces with linguistics and the engineering sciences may be jointly directed to the investigation of many particular applications. One socially important application is that of the medical aspects of speech. In this thesis, phonetics and signal processing concepts and methods are combined to offer the medical sciences of laryngology and speech pathology an objective quantification of laryngeal performance in speech. The medical sciences should benefit from research into a system which may be

applied to the screening and diagnosis of laryngeal disorders, while the speech sciences will benefit from the unification of theories describing the normal function of speech and the pathological malfunction of speech.

Laryngologists, speech pathologists and audiologists are primarily concerned with the detection and diagnosis of communication disorders which enable prompt treatment and rehabilitation. For the perception side of the communication system, audiologists have a large battery of audiometric tests at their disposal for the evaluation of the peripheral and central components of the auditory system -- these tests are often used in schools and clinics as part of a general program of screening for a variety of pathologies (e.g. visual, articulatory, cardiac, blood, etc.). Speech pathologists are responsible for the assessment of the production side of the communication system which includes the phonatory aspects of speech production. There is, however, no comparable situation in the speech production area. At present, a straightforward battery of tests is not generally available to laryngologists and speech pathologists for the quantification of deviant laryngeal behavior associated with pathology of the larynx. This thesis is an attempt to remedy this situation to a certain degree through the development of an acoustic system for the analysis of perturbations found in speech signals produced by pathological as well as healthy speakers.

The chief concerns of the laryngologist are medical diagnosis and treatment of the pathologies of the larynx. Two of the basic techniques used by the laryngologist to diagnose the various disorders of the laryngeal mechanism are 1) visual examination of

the larynx and 2) auditory evaluation of the voice. Visual laryngeal examination is usually completed by indirect laryngoscopy in which a mirror is placed in the patient's throat in order to observe the vocal folds and surrounding tissues. This visual assessment technique provides a restricted supralaryngeal view of the larynx under static conditions. Other instruments such as the direct laryngoscope can be used for improved visual examination of the larynx but all these techniques are intrusive and may not be readily accepted by certain patients. Both the indirect and the direct methods may be combined with the technique of stroboscopic illumination, to provide a more dynamic view of vocal fold activity.

Auditory assessment of the phonatory quality of a patient's voice suffers from extraneous factors, with differences between individuals' perceptions being the greatest compounding factor. For instance, Perkins (1977) found 27 terms in the relevant literature for describing defective voices. It was noted that disagreements between listeners even occurred for the variety of vocal qualities produced by speakers with unimpaired laryngeal mechanisms. Recent research into the phonetic description of voice quality may resolve some of these perceptual disagreements as well as standardize the descriptive vocabulary (see, for example, Laver 1980; Laver and Hanson 1981; Laver, Wirz, Mackenzie and Hiller 1981). In the present study, the intention is to develop an acoustic technique for describing and evaluating laryngeal function which is complementary to the ones already used by the laryngologist. The benefit of an acoustic technique is that it is non-invasive, and produces objective quantitative measures of dynamic laryngeal behavior.

Speech pathologists are responsible for the behavioral rehabilitation of laryngeal function following medical diagnosis and treatment of laryngeal pathology. Assessment of rehabilitative progress currently tends to rely primarily on subjective perceptual evaluations of phonatory voice quality. Therefore, speech pathologists would particularly benefit from an evaluation technique which provides objective quantitative measures which complement the subjective evaluation of a patient's phonatory behavior. Indeed, the patient, laryngologist and speech pathologist could all benefit from the improved means of communicating diagnostic information which a detailed, objective, permanent record of laryngeal behavior can help to support. The development of a quantitative objective assessment of laryngeal function would supply the speech pathologist with an additional screening tool to be used in schools and clinics alongside tests of articulation and language usage. Most of the patients examined by the laryngologist are self-selected ones who often display laryngeal pathology in its more advanced stage of development. The importance of early detection of laryngeal pathology cannot be overstated. For example, Bryce (1974) noted that for many patients, cancer of the larynx had been present and producing symptoms for an average of 6 months before a correct diagnosis had been made. Non-invasive screening of selected populations by the speech pathologist with a quantitative assessment tool could detect persons who require the immediate attention of the laryngologist.

The basic interest of the present study for speech scientists is, as noted, the unification of the normal model of speech production with the non-normal aspects of speech. The ability to

quantify the laryngeal behavior of healthy and pathological speakers will additionally be useful in a number of areas covered by the speech sciences. Firstly, successful speaker characterization of the phonatory aspects of speech has direct implications for the development of systems for speaker recognition. Secondly, the ability to characterize speakers is the necessary first step in the construction of speech recognition systems which must be designed to minimize non-linguistic information contained within the speech signal. Thirdly, the quantification of dynamic laryngeal behavior provides useful parametric modeling information for speech synthesis systems which are, at present, generally intelligible but lack the naturalness of human voice quality.

Why choose an acoustic analysis technique when other methods are currently widely available for quantitative laryngeal research and diagnosis (e.g. electrolaryngography, glottography, high-speed cinematography, stroboscopy, pneumotachography, electromyography and laryngoscopy)? The outstanding characteristic of acoustic analysis is its non-invasive nature which makes it suitable for routine clinical assessment of laryngeal function and, in particular, for screening of the general population (or more limited populations) for the presence of laryngeal pathologies. If voice samples are collected by standard tape recording procedures then speakers experience minimal distress and the system becomes highly portable. However, the validity of acoustic assessment procedures is dependent on the complex relationship between the vibrating source function and the resultant speech signal output by the production system. The nature of this complex relationship can be summarized from Davis (1979:274) as follows:

1) In general, asymmetrical changes in the mass and elastic properties of the vocal folds are created by the presence of laryngeal pathology.

2) These asymmetrical changes result in the modulation of the subglottal airstream by unbalanced vocal fold movement.

3) Irregular air pulses are emitted by the larynx into the supraglottal structures which are then radiated at the lips and nose.

4) The resultant acoustic signal is therefore affected by a disturbance of the vocal folds -- and the acoustic speech signal can be used to quantify the disturbance.

Noting the complexity of this relationship, Koike, Takahashi and Calcaterra (1977) suggested that the wide variability of acoustic analysis results which has been reported in early investigations is due to the difficult task of separating source effects from supraglottal filter effects.

With the advent of more advanced analysis techniques (in particular, digital signal processing of acoustic waveforms), recent investigations have begun to produce more precise quantitative acoustic data on healthy and pathological speakers. Most acoustic analyses of healthy and pathological laryngeal function have focused on either spectral aspects (e.g. interharmonic energy and spectral slope) or on the periodic attributes of the speech waveform. The emphasis of this thesis will be on the assessment of laryngeal function via the acoustic analysis of irregularities in the periodic structure of speech signals. The main aim of the study is the development of a computer-based system which can differentiate between groups of healthy and pathological speakers by means of an acoustic analysis of speech signal perturbations of fundamental frequency and amplitude. The successful development of such a system is the first step towards its use as a screening tool for

detecting the presence of laryngeal pathologies in the general population.

For the present study, the area of interest has been restricted to the development of an acoustic analysis system for evaluating fundamental frequency (F0) characteristics evidenced in the phonations of healthy and pathological speakers. This aspect of phonation has been the focus of a major part of the research into the acoustic characteristics of laryngeal pathology. The fundamental frequency parameter can be acquired by a variety of acoustic analysis techniques -- there is a considerable amount of signal processing research in this area due to the relevance of the F0 parameter, noted above, to a number of tasks such as speaker and speech recognition, synthesis, vocoders, etc. On close inspection, the time domain speech signal reveals two important characteristics of fundamental frequency. The repetition of pitch periods in the voiced segments of the signal, even in normal, healthy phonation, does not display a perfectly smooth-changing sequence of durational values. The very short-term duration of each successive cycle tends to vary, in a somewhat random manner, from the longer-term general trend which describes the overall movement of the periods. A number of terms related to these two F0 characteristics will be operationally defined as follows:

1) An acoustic analysis technique will be used to extract individual period duration values from a sample of connected speech -- these period measurements will be converted to F0 values in units of Hz for ease of presentation. A series of F0 values extracted from an input sample of speech will be termed an F0 contour. In

addition, the algorithms to be used to measure periodic activity in speech waveforms will also extract peak amplitude (A0) values associated with each cycle of vibration. A series of A0 values extracted from a speech waveform will be termed an A0 contour. F0 and A0 contours should display short-term random movements of their individual values about the long-term movements of the contours.

2) In order to evaluate the short-term movements seen in F0 and A0 contours, a filter will be applied to these contours to smooth out the random movements of the individual values. The results of the smoothing procedures will be two new functions, namely the F0 trend line and the A0 trend line. The trend line of F0 values extracted from an entire sample of connected speech will be used to produce long-term measures (e.g. mean and standard deviation) which will often be called "intonational" parameters. Here, one needs to distinguish between this restricted acoustic meaning and the broader usage of the term "intonation" in the linguistic and phonetic literature.

3) The local individual deviations of F0 contour values from the equivalent smooth F0 trend line values will be termed frequency perturbations. Long-term parameters of frequency perturbation will also be derived from an entire sample of connected speech -- perturbation analysis has been the classic technique for evaluating pathological conditions of the vocal folds since these disturbances are perceived in terms of an auditorily "rough" phonatory quality. The increased disturbance of the anatomy and physiology of the vocal folds should produce increased degrees of

frequency perturbation as pathology develops, with an associated greater degree of perceived "roughness" (Hiller, Laver and Mackenzie 1983). In the literature, frequency perturbation parameters are often referred to as pitch or frequency "jitter". As in the case of F0 measurement, a smooth trend line of A0 values will be determined for a sample of connected speech and individual deviations of A0 measures evaluated for amplitude perturbations. Amplitude perturbations are often called "shimmer" in the relevant acoustic studies.

SECTION 1.1 — OUTLINE OF THE ACOUSTIC SYSTEM

The following outline is a brief introduction to the acoustic measurement system to be discussed in detail in the remaining chapters of this thesis.

The automatic system for the acoustic analysis of F0 and A0 perturbations in connected speech consists of three major stages. In the first stage, an input speech waveform is analyzed for F0 and A0 data by a parallel processing pitch detection algorithm operating in the time domain, devised originally by Gold and Rabiner (1969). The outputs of the pitch detection algorithm are two contours, one consisting of F0 values while the other consists of amplitude values which are based on peak measures in the speech waveform. In the second stage, a non-linear smoother, based on the work of Rabiner, Sambur and Schmidt (1975), is applied to each contour to produce trend lines which preserve long-term movements but with the rough components associated with the individual deviations of the F0 and A0 values smoothed out. These deviations of F0 and A0 from their trend lines, to be called excursions, are the basic units for

measuring perturbations of frequency and amplitude. The use of the trend lines limits the effects of long-term movements of F0 and A0 from the analysis of perturbations in connected speech. In the final stage of the analysis system, the excursions of F0 and A0 values from their associated trend lines are statistically evaluated for perturbation parameters. Most of these measures are distributional in nature, providing such parameters as the mean and ranges of the waveform jitter and shimmer. Other special parameters of waveform perturbation are also evaluated at this stage. The general flowchart of the perturbation analysis system is presented in Figure 1.1.

SECTION 1.2 — OUTLINE OF THE REMAINING CHAPTERS

In Chapter 2, a typology will be presented for the wide variety of pitch detection algorithms which have been reported in the speech signal processing literature. This typology is intended as a framework in which the pitch detection algorithm selected for implementation in the present study of laryngeal function may be related to other available types of pitch detector.

The implementation of a time domain pitch detection algorithm will be presented in full detail in Chapter 3. This time domain algorithm is a modified and elaborated version of the parallel processor developed by Gold and Rabiner (1969). The parallel processor is considered appropriate for perturbation analysis since it uses a multichannel approach to pitch detection which can operate on a wide variety of temporal structures in the speech waveform, over a fundamental frequency range of approximately 50 to 600 Hz. Several issues related to the implementation of the parallel

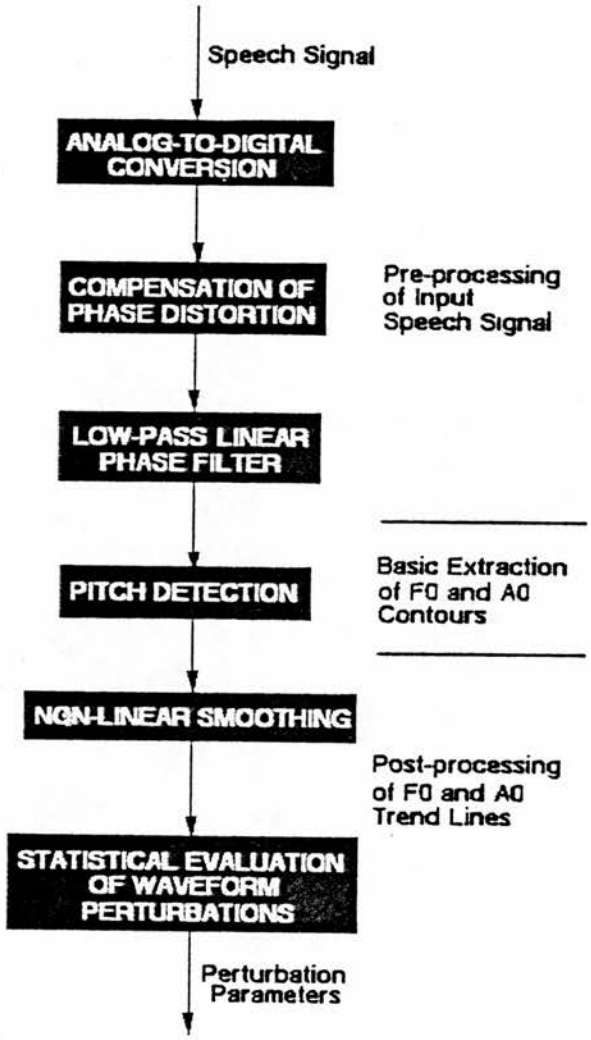


Figure 1.1 General flowchart of the perturbation analysis system to be discussed in this thesis.

processing detector will also be discussed in Chapter 3. Firstly, the appropriate analysis conditions for obtaining useful results from the parallel processor are to be considered. Secondly, the performance of the modified parallel processor will be compared to visual examination of samples of connected speech. Thirdly, the issue of the sampling resolution used to first quantize the speech signal will be discussed in relation to the resolution of the pitch detection results. The chapter will conclude with a discussion of the range of applications for which the parallel processor is suited.

The parallel processing pitch detection system produces F0 and A0 contours which can be analyzed for perturbatory behavior. A review of the literature relevant to perturbation analysis will be set out in Chapter 4. Well over 30 studies have been published in this literature in which pitch period, F0 and A0 contours have been examined for perturbatory behavior, in particular, as a method for quantifying laryngeal function associated with voice pathology. This literature review is intended as a framework in which the present study may be related to the work of other researchers. In particular, it will be seen that the analysis of connected speech samples for F0 and A0 perturbations using a trend line approach has not yet received much attention in the relevant fields.

The implementation of the algorithms used for evaluating F0 and A0 perturbations will be described in detail in Chapter 5. The measurement of F0 and A0 perturbations will begin with the derivation of smoothed trend lines by the application of a non-linear smoother to the input contours. This smoother will be a digital system consisting of a 5-point median filter and a 3-point

Hanning window, first described by Rabiner et al. (1975). The construction of the smoothed trend lines will permit the measurement of deviations between the input contours and their smoothed equivalents -- these deviations will be the basic units of perturbation known as excursions. A number of long-term measures of F0 and A0 perturbations are based on the excursion values including distributional measures such as the range and percentage of substantial excursions of both F0 and A0. This chapter will conclude with a short section on the basic procedures for deriving perturbation and intonation parameters from samples of connected speech.

In Chapter 6, the perturbation measurement system will be applied to a large number of voice samples produced by healthy and pathological speakers in order to evaluate the system's usefulness for the detection of laryngeal pathology. This evaluation is a necessary precursor to any potential application of the perturbation measurement system as a screening tool for detecting the presence of laryngeal pathology in the general population. Six sets of experiments will be reported in this chapter -- the first 2 experiments are concerned with the requirements of the speech sample to be input to the perturbation measurement system while the remaining 4 experiments evaluate the ability of the system to differentiate between groups of healthy and pathological speakers.

In summary, the overall objective of this dissertation is to develop and assemble a robust and accurate acoustic pitch detection system, together with a method for analyzing waveform perturbations, and apply it to the recorded speech of groups of speakers with known laryngeal pathology. The results will be compared to those for a

large control group of speakers, thought to be healthy, and the feasibility of using the system for screening, monitoring and diagnosis of laryngeal disease will be explored.

CHAPTER 2

PITCH DETECTION ALGORITHM TYPOLOGY

CHAPTER 2

PITCH DETECTION ALGORITHM TYPOLOGY

2.0 INTRODUCTION

The primary goal of the present study is the development of an automatic system for the acoustic analysis of waveform perturbations evidenced in speech samples produced by healthy and pathological speakers which will be useful for the screening, differential diagnosis and rehabilitation of laryngeal pathologies. An integral part of this automatic system is the pitch detection algorithm (PDA) which extracts the initial fundamental frequency and amplitude data from samples of connected speech, this information being the primary input to a set of perturbation measurement algorithms. The choice of a PDA is critical to the aforementioned development goals, particularly since many algorithms designed for use with the acoustic characteristics of normally-produced speech cannot be effectively applied to the perturbed signals of dysphonic speech. In this section, a typology of PDAs is presented as a framework from which a useful detector can be selected from the many algorithms presented in the literature, this PDA being appropriate to the measurement of frequency and amplitude perturbations evidenced in speech waveforms. The typology presented in this chapter is not an exhaustive account of all pitch detection algorithms which have appeared in the literature. The focus is on the more well-known PDAs which have been investigated in the speech signal processing area. A very complete coverage of pitch detection systems can be found in the book by Hess (1983) which, together with the

classification work of McKinney (1965) and Rabiner and Schafer (1978), has strongly influenced the PDA typology presented in this thesis.

The following discussion begins by categorizing all PDAs into two domains -- a given PDA may operate primarily in the time domain or in the spectral domain. Though no single PDA is capable of accurately analyzing the complete range of signals produced by the speech mechanism, this review of the pitch detection literature demonstrates the greater usefulness of a time domain based multichannel solution to the pitch extraction task as applied to waveform perturbation analysis.

A number of relevant points will be mentioned prior to a complete discussion of the PDA typology. Firstly, the term "pitch detection" as it is applied here is meant to be a general one which covers the various aspects of signal processing used to derive acoustic period or fundamental frequency parameters from a given speech waveform (Hess 1983). Though this term is primarily associated in the psychoacoustic literature with the interaction of physical stimulus and mental perception, the labels "pitch" and "pitch detection" have been widely used in the signal processing literature in a broader sense, to include the measurement of signals for acoustic, perceptual and linguistic phenomena. As this thesis is set in a speech signal processing context, the broader use of the terms "pitch" and "pitch detection" has been adopted. Specific reference to the psychological phenomenon of perceived pitch (as opposed to the acoustic measurement of pitch parameters) will be made where appropriate to the discussion.

Secondly, emphasis will be placed on pitch detection algorithms which automatically extract pitch data from the acoustic speech signal. This thesis is primarily concerned with time and spectral domain PDAs. Other areas related to the pitch detection field will only be mentioned here and not discussed in detail. These related areas include: 1) manual pitch detection (e.g. the visual examination of spectrograms and oscillograms); 2) pitch detection instruments including optical (e.g. high-speed photography), and electrical (e.g. laryngography) and mechanical (e.g. Sondhi tube for inverse filtering) devices; 3) manual voicing detection methods (as in 1) above); 4) automatic voicing detection algorithms. It should be noted that the outputs of some of the pitch detection instruments can be analyzed by some of the algorithms to be discussed in the following sections (e.g. time domain algorithms can be applied to the output of a laryngograph or throat microphone). No detailed discussion is presented for voicing detection algorithms other than to point out that some PDAs require separate algorithms for voicing detection while others contain logic which satisfies both pitch extraction and voicing detection needs.

Thirdly, it is relevant to describe the characteristics of the acoustic material to be analyzed by a pitch detector in the present study. Most PDAs are designed with an idealized notion of the acoustic material to be examined for pitch data. This ideal data is assumed to be produced by a young adult male with optimally efficient laryngeal vibrational characteristics in an acoustic environment characterized by a good signal-to-noise ratio. This type of efficient vibration has a number of acoustic and physiological conditions associated with it (Laver 1980; Laver and

Hanson 1981; Laver, Hiller and Hanson 1982) including:

1) Physiological Factors
   a) Regularly periodic vibration.
   b) The true vocal folds are the sole source of the vibration.
   c) Moderate settings of all myodynamic parameters are used to produce full glottal vibration.
   d) Air is used in an efficient manner without the production of audible glottal friction.

The resultant laryngeal pulse shape for this type of vibration is approximately triangular with the maximum excitation point occurring during the closing phase of the glottal cycle; this closing phase lasts for about 33% of the cycle.

2) Acoustic Factors
   a) A glottal waveform showing certain spectral characteristics including a harmonic slope of -10 dB/octave below 250 Hz and -12 dB/octave above it.
   b) Average fundamental frequencies which range from 45 to 250 Hz.
   c) The larynx pulse sequence demonstrates a limited range of jitter and shimmer, these waveform perturbations being normally distributed with a standard deviation of 2% or less of the mean fundamental frequency and amplitude (Hanson 1978).

In reality, many if not all of the above ideal characteristics are disturbed by dysphonic conditions of the voice, thus presenting significant problems for automatic pitch detection. Two major sources of difficulty found in dysphonic voices (Laver 1980; Laver and Hanson 1981; Laver et al. 1982) include:

1) The presence of fricative, non-harmonic energy in the speech waveform. Perceptually, this additional friction gives a "whispery" effect to phonatory quality.

2) Frequent and substantial perturbations of period and amplitude which produce excessive jitter and shimmer. Jitter and shimmer may be evidenced in two ways including:
   a) In a quasi-random production of laryngeal vibration which produces a "harsh" auditory voice quality and
   b) In a pulse-grouping tendency (e.g. a sequence of long-short-long-short cycle durations) which results in a "creaky" auditory effect.

Whispery, harsh and creaky phonation types are common features of the speech of dysphonic speakers. Speakers with normal phonation mechanisms also depart from the ideal phonatory conditions -- the majority of speakers in the general population show evidence of varying degrees of phonatory inefficiency, in particular, whispery and creaky phonation types. In addition, the recording environments found in many clinics do not demonstrate good signal-to-noise conditions for the collection of voice data.

In setting out a typology of PDAs, it will be useful to describe a number of criteria which a PDA can be compared against in order to determine its usefulness in the analysis of the perturbed waveforms produced by pathological voices. In stating these criteria, it is not suggested that a single PDA presently exists which fulfills all the requirements. However, given the great variety of pitch detection schemes which are present in the literature, a system or systems should exist which offers at least a partial solution to the pitch detection problem as applied to dysphonic speech. The following characteristics should be demonstrated by a PDA for adequate performance on dysphonic speech data (Laver et al. 1982). It should:

1) Work on acoustic recordings of the wide range of fundamental frequency produced by men, women and children

2) Be noise resistant
   a) be resistant to the effects of non-harmonic noise from laryngeal friction
   b) be accurate to within 2% in tracking F0
   c) retain accuracy over wide F0 and jitter and shimmer ranges
   d) be relatively impervious to poor signal-to-noise ratios arising from poor quality clinical recordings

3. Work on continuous speech
   a) have an adequate voicing detector
   b) use a trend line approach to cope with the more

long-term movements of F0 contours, and to provide a baseline from which to measure perturbatory excursions of individual periods.

In the following sections, a number of issues relevant to automatic pitch detection are discussed in detail. The discussion begins with the basic division of all PDAs into those systems which operate primarily in the time domain or in the spectral domain. Within each domain of operation, the structure of the input speech waveform as well as the techniques for deriving pitch data from these signals are examined. In terms of signal structure, relevant issues include waveform resolution, the effects of spurious noise, phase distortion and the computational complexity of the analysis techniques. A great number of PDAs exist in the literature and the general typological approach of Hess (1983) is used to highlight certain trends for pitch analysis within the two domains of operation. For time domain PDAs, various temporal structures within the time domain waveform are emphasized by pre-processing of the input signal prior to the extraction of pitch periods. For spectral domain PDAs, a number of spectral transformation techniques are capable of emphasizing spectral harmonic components in a manner which facilitates the extraction of pitch period or fundamental frequency data.

Time Domain and Spectral Domain Pitch Detection Algorithms

The basic division of PDAs into time and spectral domains begins with a description of the general realization of these algorithms. For a more detailed discussion of some of these algorithms, the reader is referred to the original accounts, and to Hess (1980; 1982; 1983) and McKinney (1965). Figure 2.1 (after

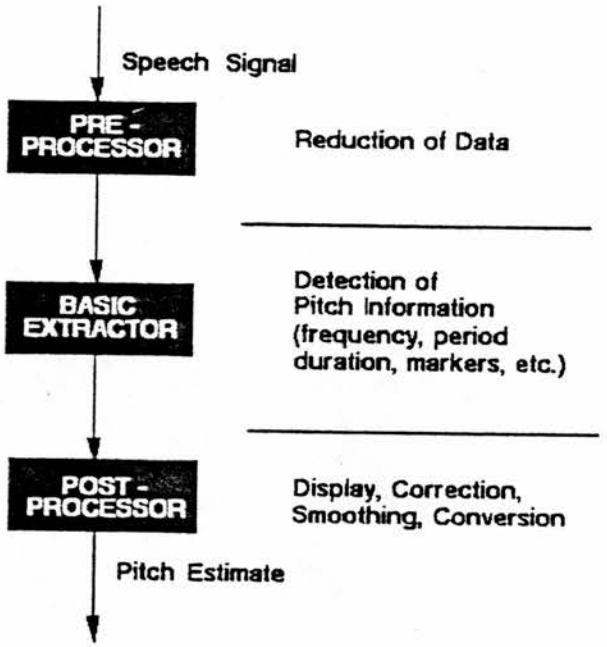


Figure 2.1 Block diagram of the general realization of algorithms for pitch detection. (After McKinney 1965 and Hess 1983).

McKinney 1965) displays the typical components of a PDA through which a signal passes, including the pre-processor, basic extractor and post-processor. The pre-processor applies some form of data reduction to the input speech signal in order to simplify the pitch extraction task. The main task of pitch estimation is then achieved by the basic extractor which converts a pre-processed signal into a series of pitch estimates. The post-processor can be used for a variety of tasks depending on the nature of the PDA and its particular application. For example, the post-processor may be used for error detection and correction, smoothing of pitch contours and the display of pitch parameters. Given this general realization, the categorization of PDAs is based on the domain of operation of each PDA, which is defined as the domain of the signal which is input to the basic extractor component of the detector system (Hess 1983). This definition results in the subdivision of PDAs into time domain and spectral domain PDAs (see Figure 2.2). To be classified as a time domain system, the basic extractor of a PDA extracts pitch data from a signal whose time-base is equivalent to the time-base of the original input speech waveform. The remaining algorithms are classified as spectral domain extractors. The primary characteristic of this type of algorithm is the short-term transformation of the speech signal performed at the pre-processing stage. A given short-term transformation converts the input time domain signal to some spectral domain. This category includes those algorithms usually labeled as frequency domain (e.g. cepstral pitch extraction) as well as transformations into the lag domain (e.g. autocorrelational analysis). The short-term analysis characteristic resolves, to a certain degree, the confusion which is often associated with the labeling of lag domain algorithms as either time

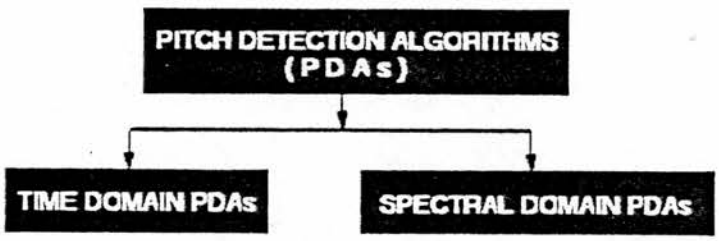


Figure 2.2 Primary subdivision of pitch detection algorithms.

or frequency domain. Other researchers such as Ungeheuer (1963), McKinney (1965) and Rabiner, Cheng, Rosenberg and McGonegal (1976) have distinguished between PDAs which operate in the time domain or the frequency domain. Hess (1983) uses the label short-term analysis to classify techniques as spectral domain as opposed to time domain methods. It should be noted that the exact distribution of particular algorithms into these various categories differs in each of the aforementioned studies by Hess and McKinney.

The output of the basic extractor depends on the domain of operation. Time domain algorithms produce indicators of laryngeal pulse behavior. These indicators are often termed pitch markers as they mark the boundaries of the measured periods within the time domain signal. Thus, the analysis of pitch is a local one for time domain algorithms since the actual location of each period is determined from the time domain waveform. Spectral domain algorithms produce an inherently smoothed estimate of period or fundamental frequency for a given frame of data. This is a global analysis of period or F0 since a single estimate of pitch is derived for a given frame of speech containing a small number of individual periods.

SECTION 2.1 — TIME DOMAIN PITCH DETECTION ALGORITHMS

General Realization of Time Domain PDAs

The oldest methods of automatic pitch period extraction operate on the time domain representation of the speech signal. A typical time domain extractor as described by McKinney (1965) is displayed in the block diagram of Figure 2.3 (after Hess 1983:152). Most time domain

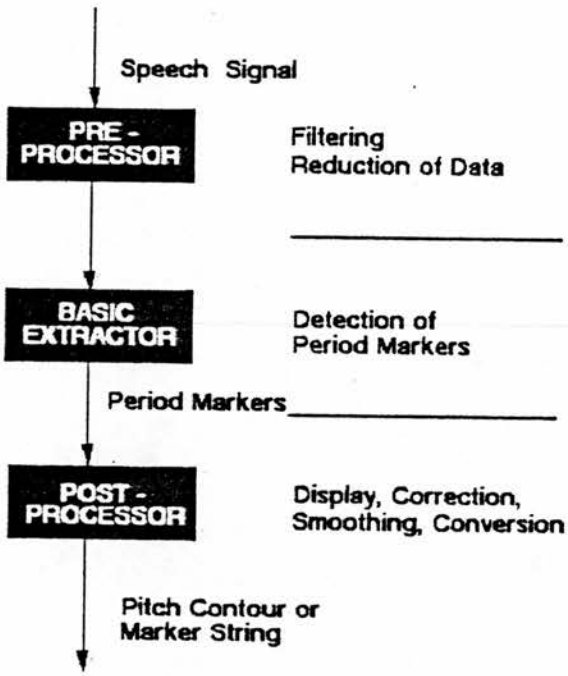


Figure 2.3 Block diagram of a typical time domain pitch detection algorithm. (After Hess 1983:152).

extractors include a pre-processor, basic extractor and post-processor in their design. The task of the pre-processor is data reduction in which the speech signal is filtered by linear and/or non-linear functions. The pre-processor simplifies the pitch extraction task by removing undesirable components from the signal (e.g. formant frequencies which may reinforce non-fundamental harmonics) and emphasizing the temporal structure of interest within the waveform. The filtered speech is input to the basic extractor to determine the locations of the pitch markers which indicate the laryngeal pulse behavior evidenced in voiced segments of speech. The output of the basic extractor is a sequence of markers, each marker being associated with a pre-defined significant point within each period (e.g. peaks or zero-crossings). The post-processing technique is dependent on the particular application for which pitch period data is required. In particular, the post-processor is required to correct various extraction errors produced by the basic extractors of certain time domain devices.

The Ability of Time Domain PDAs to Measure Individual Periods

The time domain extractor is, in theory, capable of determining each individual pitch period present in the speech signal. This ability to process individual periods is the main advantage of the time domain device since the basic extractor can accurately process 1) rapid increases and decreases in period duration and 2) slightly irregular signals (e.g. perturbed waveforms associated with harsh and creaky types of laryngeal voice quality). However, this ability is also the main disadvantage of time domain systems since the direct extraction of pitch markers may lead to errors when the

signal has been corrupted by noise.

Significant Periodic Features in the Time Domain Waveform

A number of significant features in the speech waveform have been noted by Hess (1983:153-154) as useful indicators of periodicity. The first two features are characteristic of any periodic or quasi-periodic signal. Firstly, there is the presence of the first partial (i.e. the fundamental harmonic) within the signal for a voiced sound which has a cycle duration equal to the fundamental period of the waveform. Secondly, a periodic signal displays an overall structural pattern which is replicated from cycle to cycle of laryngeal vibration. Two additional features of periodicity are derived from the source/filter characteristics of the speech signal. The linear model of speech production assumes that the periodic speech signal is the output of an approximately time-invariant linear filter which has been stimulated by a pulse generator. The impulse response of a linearly passive system such as the vocal tract consists of the sum of all the exponentially decaying sinusoids (Fant 1960). The third feature of periodicity is the resultant waveform structure which is characterized by high amplitudes at the beginning of each period followed by the decay to low amplitudes near the end. The final feature is based on the presence of discontinuities within the speech waveform for voiced speech. The stimulation of a linear system by a pulse train may result in discontinuities of the output signal at those instants where the individual pulses occur. If the discontinuities are not evidenced in the waveform itself, then they should be displayed in a first- or higher-order derivative of that waveform.

Requirements of the Time Domain Signal

Time domain PDAs are designed to extract periodic features directly from the speech waveform (or a low-order derivative of the signal), these features being markers of the cycles within voiced sections of speech. Time domain extractors must derive these features from many different signal types produced by speakers. In the first instance, the overall performance of a time domain PDA depends on the structural condition of the waveform itself. This structural condition may be effected by a number of things, in particular, the resolution of the signal, the addition of noise to the signal and low-frequency phase distortion caused by various recording equipment.

Sampling Resolution — Many of the time domain algorithms extract pitch data from a digital representation of the speech waveform. Thus, the measurement accuracy of a given cycle of voiced speech is directly dependent on the original sampling resolution used to quantize the signal during the analog-to-digital conversion process. For a fixed sampling resolution, the accuracy of period measurements will vary within the continuous speech of an individual speaker as well as between the speech of different speakers (this is most notable for the octave difference in F0 between most male and female voices). The degree of resolution of the time domain signal depends on the purpose for which the pitch data is being extracted. In general, Hess (1983:83-88) notes three pitch detection tasks which require differing degrees of measurement accuracy based on psychoacoustic findings. These tasks include: 1) the use of pitch data for speech synthesis in which the sensitivity to pitch resolution is high for the human auditory system, 2) the use of

pitch data to examine speech production phenomena where the performance of the production mechanism is considered to be generally poorer than the resolving power of the auditory system thus requiring a lower sampling resolution and 3) the examination of linguistic phenomena in which gross changes in perceived pitch elicit important linguistic responses thus requiring an even lower resolution in the original signal. For certain tasks, sampling rates may need to be very high (upwards of 80 KHz) in order to accurately preserve pitch information (Horii 1979). The use of high sampling rates results in increased storage and computational processing of pitch data. To avoid these large computational requirements, interpolation techniques have sometimes been used to increase the signal resolution by mathematical means. The issues of sampling resolution and interpolation are discussed in further detail in Section 3.3 below.

Noise — The extraction of periodicity features from the time domain signal will be corrupted by the addition of noise to the signal. If the noise is continuous in nature (e.g. tape recorder hiss) then most time domain PDAs will still extract accurate data if the relevant features have not been masked by the noise. A special case of additive wide-band noise is the friction produced by air flow through the glottis during the periodic phonation of some speakers which produces a "whispery" auditory sensation. Pre-filtering processes incorporated into a number of time domain PDAs help to reduce the influence of wideband noise. Narrowband periodic background signals are most problematic for time domain PDAs. In particular, the presence of a 50 or 60 Hz electrical hum will modulate the entire waveform especially for low level signals

within a given utterance. This hum can be high-pass filtered from the signal but this filtering will limit the lower end of the fundamental frequency range which can be measured by certain types of time domain PDA. Another problem is the occurrence of transient environmental noises within the utterance to be analyzed which completely disrupt the structure of the signal.

Low-frequency Phase Distortion -- The detailed structure of the time domain speech waveform will be affected by low-frequency phase distortion during the recording and playback of a speech sample. The relative phases of the harmonics of the signal are not maintained due to frequency-dependent responses of the electrical components used in the recording equipment. This phase distortion should not be severely problematic for time domain pitch detection since the fundamental periodic structure is preserved in the recorded waveform. However, it still may be useful to prevent or correct phase distortion of speech data for pitch analysis. Firstly, standardization of speech materials is required if accurate comparisons are to be made of speech analysis results produced within and between laboratories. This standardization could be achieved by the recording of signals with known harmonic structures in order that appropriate phase correction procedures may be completed. Secondly, many time domain PDAs examine the speech waveform for such features as peak minima and maxima and zero-crossings which act as anchor points from which periodicity may be derived -- these anchor points may be laterally shifted in apparent time by low-frequency phase distortion. It may be useful to maintain or restore the original relationships of these features within the waveform in order to improve pitch extraction results. A

method for phase compensating speech waveforms for time domain pitch extraction is discussed in Section 6.2 below.

The General Modes of Pitch Detection in the Time Domain

The various features of periodicity evidenced in speech waveforms have broadly determined the modes of operation for time domain PDAs. Based on these waveform features, time domain PDAs may be summarized from Hess (1983) as operating in the following manners:

1) Time domain PDAs which process the fundamental harmonic component of the speech waveform in order to determine pitch data. This type of PDA emphasizes the first partial of the signal by a substantial amount of non-linear and/or linear pre-processing of the input signal. The emphasized fundamental harmonic is then analyzed by relatively simple threshold techniques in the basic extractor which produce the pitch markers.

2) Time domain PDAs which process the overall temporal structure of the signal. In structural analysis, the basic extractor operates directly on the speech signal and therefore must cope with a variety of signal structures. If any pre-processing is completed, it is usually in the form of a moderate low-pass filter to eliminate some high-frequency components from the signal.

3) Time domain PDAs which process a time domain signal derived from the original signal -- this derived signal is related to the source excitation waveform as described in the linear model of speech production. This structural simplification technique incorporates elaborate pre-processing techniques to produce the derived excitation signal which is then examined by structural analysis techniques to extract periodicity information.

A number of multichannel approaches which incorporate various features from these operating modes have been applied to time domain analysis of speech in an attempt to improve pitch extraction results. The following sections discuss these general approaches to pitch extraction in the time domain.

SECTION 2.1.1 -- TIME DOMAIN PDAS WHICH PROCESS THE FUNDAMENTAL HARMONIC COMPONENT OF THE SPEECH SIGNAL

For a voiced segment of speech, it can often be observed that a first partial (i.e. the fundamental harmonic) is present at a period equal to the pitch period of speech. This observation is the basis of one type of time domain detector which directly extracts the fundamental harmonic from the speech signal. This type of device is usually comprised of an elaborate pre-processing stage followed by a relatively simple basic extraction of the fundamental harmonic. The pre-processor is designed to produce a new time domain waveform in which the fundamental harmonic has been emphasized relative to the other components in the waveform. The basic extractor is an event detector which determines the locations in time of the fundamental harmonic characteristics emphasized by the pre-processor. The output of the basic extractor is a sequence of pitch periods derived from the detected events of the fundamental harmonic. The success of the fundamental harmonic extractor is dependent on two assumed characteristics of the input speech waveform. Firstly, the fundamental harmonic must be present in the speech waveform. Therefore, the fundamental harmonic extractor cannot be applied to bandlimited signals, for example, telephone transmitted speech in which the first partial has been filtered out. Secondly, the input speech waveform should not be distorted in the low frequencies which would result in corruption of the fundamental harmonic. Often called frequency meters, fundamental harmonic extractors are the classic systems built in analog hardware. In fact, McKinney (1965) classified fundamental harmonic devices as frequency domain techniques since the basic extractor measures the

frequency of occurrence of certain characteristics related to the first partial. However, frequency meters are actually time domain devices since a measure of pitch period is derived from an examination of the time domain waveform (Hess 1983). The following sections summarize the reviews of McKinney (1965) and Hess (1983) on the pre-processing and basic extraction aspects of fundamental harmonic extraction.

Hess (1983:156) subclassifies the fundamental harmonic detector by the three types of basic extractor often found in these devices including 1) the zero-crossings analysis basic extractor (ZXABE), 2) the threshold analysis basic extractor (TABE) and 3) the TABE with hysteresis. The ZXABE detects an event each time the input signal crosses the zero axis with a pre-defined polarity. An event is detected by the TABE each time the signal crosses a pre-determined threshold level with a desired polarity. The addition of hysteresis to the TABE means that an event is detected when two thresholds are crossed successively in a defined sequence. The TABE with hysteresis provides an element of security during pitch extraction since an output will only be produced when both thresholds are crossed. In either TABE system, the non-zero thresholds must be normalized in relation to the signal amplitude. The output of the extractors is non-linear since each event detected is represented as an impulse of constant shape and polarity. The post-processor examines the duration between impulses to produce pitch periods or fundamental frequency estimates.

The success of fundamental harmonic extraction relies on elaborate pre-processing to emphasize the first harmonic component in the speech waveform. This first partial must be emphasized in such a way that two and only two crossings per cycle occur during the basic extraction stage. Pre-processing techniques to highlight the fundamental harmonic component use linear and/or non-linear methods to transform the input signal. Linear methods use filters to isolate the fundamental harmonic component. McKinney (1965:69) described the most appropriate filter characteristics for pre-processing as follows: 1) sharp high-pass attenuation with a cutoff frequency below the F0 range of interest to eliminate spurious low-frequency signals from the speech waveform, 2) slowly increasing attenuation in the frequency range of interest (this type of attenuation is used to de-emphasize harmonics at the high end of the permitted frequency range when a first harmonic exists near the lower end) and 3) sharp low-pass attenuation at a cutoff frequency above the F0 range of interest to eliminate all higher harmonics from the signal. In the simplest case, a fundamental harmonic detector consists of a low-pass filter followed by a ZXABE, though this method is not very successful. The design of the more useful filter characteristics is relatively straightforward for a restricted range of F0 (e.g. one to two octaves) since a low number of harmonics are present -- this restricted range requires from 6 to 12 dB/octave attenuation within the passband. The task of creating appropriate filters becomes more difficult if the detector is to be used on a wider range of frequencies, say, a 3 octave bandwidth. Here, the increased number of harmonics within the passband requires up to 18 dB/octave attenuation. This type of filter is not realizable due to dynamic range considerations. Firstly, a filter

with a 18 dB/octave attentuation applied to a 3 octave range would require a dynamic range of over 50 dB. In addition, the filter must also accommodate the intrinsic dynamic range of the input signal itself (e.g. 30 dB). Such a filter, if it were possible to construct one, would be very sensitive to spurious low-frequency signals. The alternatives to such a filter are 1) divide the signal into narrow ranges of F0 by filtering -- this may be obtained by manual pre-setting or self-adjusting tunable filters or 2) compress the dynamic range of the input signal. It should be noted that in the case of female voices, linear filtering may be all that is required for fundamental harmonic extraction since the filtered speech will often be sinusoidal in appearance particularly when the first partial is emphasized by the interaction with the first formant.

Non-linear techniques are used to emphasize the first harmonic as the sum of the difference tones of the adjacent higher harmonics. In addition, these non-linear transformations may provide additional spectral flattening of the formants. Figure 2.4 (after Hess 1983:168) displays a general block diagram of a fundamental harmonic detection system which includes non-linear processing. Optional filtering is applied to the input speech signal to eliminate higher harmonics outside the F0 range of interest -- this filtering may be more moderate compared to a system which only uses linear filtering. The non-linear transformation is then applied to the signal. Additional high-pass filtering is required for those non-linear functions which produce a DC signal upon output -- the high-pass filtering suppresses the zero frequency DC component thus re-creating an AC signal. The resultant signal is then processed by

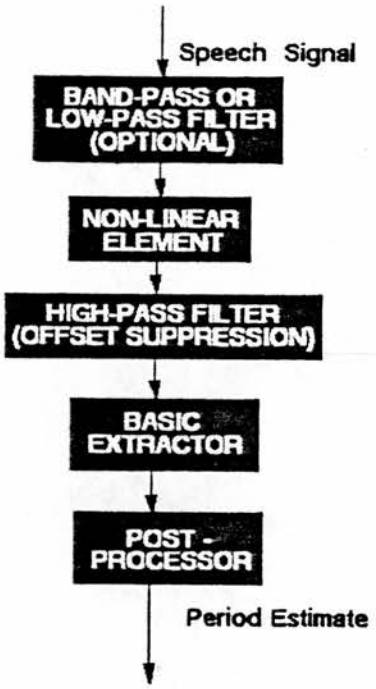


Figure 2.4 Block diagram of a time domain pitch detection algorithm based on fundamental harmonic extraction. The pre-processing stage consists of optional linear filtering of the input signal, non-linear processing and optional supression of the DC signals created by certain types of non-linear processing. (After Hess 1983: 168).

one of the simple threshold basic extraction techniques. A number of non-linear functions (usually classified by their mathematical identity) have been used in fundamental harmonic systems, each demonstrating varied success at pitch extraction. Even functions (i.e. functions which produce positive output even when the input is negative) incorporated into fundamental harmonic detectors include squaring and full-wave rectification (i.e. the absolute values of the inputs). These even functions enhance the first partial when it has a weak presence in the signal. However, the even functions produce detrimental effects when the signal contains a strong first harmonic. For example, speech waveforms produced by female speakers often are sinusoidal in appearance such that full-wave rectification produces a frequency doubling of the first partial. Odd functions are those transformations which produce a negative output when the sign of the input is negative. Odd functions used for fundamental harmonic detection include the identity function (i.e. the actual speech signal) and logarithmic functions (a form of signal compression). Odd functions have been shown to be of little use when the amplitude of the first harmonic is relatively low. Mixed function such as half-wave rectification (i.e. all values of a given polarity are kept and the remaining values set to zero) produce results which are a compromise but not optimal. More research into non-linear processing for fundamental harmonic detection is required since it appears that no single non-linear function is suitable for a wide range of fundamental frequencies and signal types.

A number of extraction errors are due to the simplistic nature of the fundamental harmonic event detector, in particular, gross errors resulting from inadequate pre-processing of the input speech signal and fine errors associated with the exact location of fundamental harmonic events within the waveform. Fundamental harmonic extractors often produce gross errors by latching onto non-F0 harmonics present in the processed speech waveform. In this case, inadequate pre-processing of the input signal has not limited the signal structure to two and only two threshold crossings per cycle of voiced speech. For the ZXABE, McKinney (1965:173) presented a formula which determines the lower bound in signal energy which a pre-processor should produce in order that only the first partial is present in the input to the basic extractor. The pre-processed speech reaches the lower bound when its first partial is greater in amplitude than the summation of all the higher harmonic amplitudes multiplied by their respective harmonic number:

$$H_1 > \sum_{n=2}^{K} n \cdot H_n$$

The formulation of a lower bound is more difficult for a TABE system since the non-zero threshold must be determined in relation to the maximum amplitude of the input speech waveform. The TABE with hysteresis is more resistant to the influence of higher harmonics due to the addition of a second threshold which must be crossed by the signal in a pre-determined manner. Several types of fine extraction error have been found for the fundamental harmonic detector, most of which are related to the precise location of detected events within the waveform. One fine error is due to the

nature of the basic extractor used in the fundamental harmonic system. Location of an event by the ZXABE is straightforward since by definition a marker is placed at each detected zero-crossing. However, marker placement is more difficult for TABE systems since the threshold is not symmetrical with respect to the input signal. The marker could be located either at the point when the threshold is exceeded or where the signal returns below the threshold. Other fine errors produced by fundamental harmonic analysis are systematic ones due to the time-variant nature of speech signals. Firstly, the ZXABE system may be affected by the phase relationships of the higher harmonics which can shift the location of zero-crossings within the speech waveform. These crossing shifts will produce a noisy F0 signal even with a high degree of low-pass filtering at the pre-processing stage. Secondly, TABE systems are sensitive to amplitude changes of the input signal. As the input signal changes rapidly in amplitude, the relationship of the threshold crossing to the previous zero-crossing (or peak minima/maxima) also changes. The result of this amplitude variation of the input signal is a frequency modulation of the output marker sequence. The frequency modulation can be limited by locating the marker at an average point between the upward and downward crossing of the threshold. Finally, all three basic extractors may be affected by the interaction of the F0 and vocal tract resonances. During a formant transition, the interaction of the first formant with the first harmonic will result in a phase distortion of the F0. This resultant phase distortion is represented as a frequency modulation of the marker sequence -- this phase problem is greater for higher F0 levels than lower ones.

SECTION 2.1.2 -- TIME DOMAIN PDAS WHICH PROCESS THE TEMPORAL

STRUCTURE OF THE SPEECH SIGNAL

Time domain devices operating as fundamental harmonic detectors process the speech signal in a manner which enhances and isolates the first harmonic for pitch period extraction. On the other hand, Temporal Structural Analysis (TSA) systems examine the overall temporal structure of the speech wave for features which indicate the fundamental periodicity of the signal. The selection of useful periodicity features begins with the basic assumption that the speech signal is the time-variant response of the vocal tract resonance system to a pulse train produced by the vibrating larynx. The temporal structure of each F0 period represents the first part of the vocal tract's impulse response — the period is the summation of a series of exponentially damped sinusoids with the first formant usually being the dominant waveform. The structural analysis of selected waveform features is often completed using computer programs which model the visual process of pitch period extraction from an oscillographic representation of the speech signal. This type of visual pattern matching is not completely understood and therefore difficult to simulate by computer algorithms. The group of TSA detectors operate either by Envelope Modeling or Sequence of Extremes analysis. Envelope modeling uses a decaying exponential function to model the speech signal envelope in order to detect discontinuities at the beginning of each period associated with the moment of impulse response. Each new pitch period is detected when the model envelope function is exceeded by the speech signal under analysis. Thus, the major problem for envelope modeling is determining a time constant for the model decay function which matches the time constant of the speech signal's decay (primarily

due to the formants). Envelope modeling pitch period extractors are usually constructed as analog devices. Sequence of extremes extractors apply heuristic models to a direct investigation of the temporal structure of the speech signal. The waveform is searched for anchor points such as peaks from which the basic periodicity of the signal can be derived. This type of structural analysis is usually completed by computer programs. These two structural analysis methods are grouped together since they share the following characteristics (Hess 1983:182):

> '1) They start from the pitch period as the representation of the vocal tract impulse response.
>
> 2) They seldom use one single detector: extraction of periodicity is done by repetitive application of similar or identical subroutines and/or circuits, or by appropriate matching of the output of parallel processors.
>
> 3) The principal mode of operation is non-linear; the algorithmic decision statement, for instance
>
> if a less b then c; else d;
>
> where c and d stand for algorithmic statements whatsoever (sic), corresponds to the commonly used circuit in analog technology.
>
> 4) They are unbiased with respect to the errors that will occur.'

Temporal Structural Analysis by Envelope Modeling

Pitch extractors categorized as Temporal Structural Analyzers incorporate features from simplified models of speech production. One subgroup of the structural analyzers models the envelope of the speech wave to determine the fundamental periodicity of the signal. A typical example of envelope modeling for pitch detection is an analog device designed by Anderson (1960) which is a modification of an instrument by Dolansky (1955). Anderson's detector was designed

to extract pitch from a wide range of F0 (100 to 600 Hz) without the necessity of manual switching between selected pitch ranges.

In the envelope modeling approach, the voiced speech signal is the output of a simplified speech model. Each vocal fold vibration interrupts the flow of air from the lungs, resulting in the production of energy in the form of an impulse which is rich in harmonic components. Each impulse excites the vocal tract resonances which reinforce some of the harmonics while suppressing others. Upon output, each cycle of the speech wave is a truncated version of the impulse response of the vocal tract -- this cycle is the sum of several exponentially decaying sinusoids which have been simultaneously excited. The waveform within a given pitch period is characterized by a maximum amplitude and maximum slope at its commencement. To extract the pitch period, it is modeled by a non-linear process in which an exponential function with an appropriate rate of decay is applied to the input speech. The decay function begins with its maximum value at the principal peak of the period and then it decays such that the function is above the remaining components of the period. When the function intercepts the speech signal then the beginning of the next pitch period has been found.

Appropriate circuitry is required which recognizes the maximum amplitude and slope properties of the pitch period and produces one output pulse per cycle. One example of this circuit functions as a combination of peak detector and exponential decay to produce the initial envelope model of the input speech wave (see, for example, the first two rows of Figure 2.5a-d below, from Hess 1983:185). This circuit determines the location of the principal peaks of the

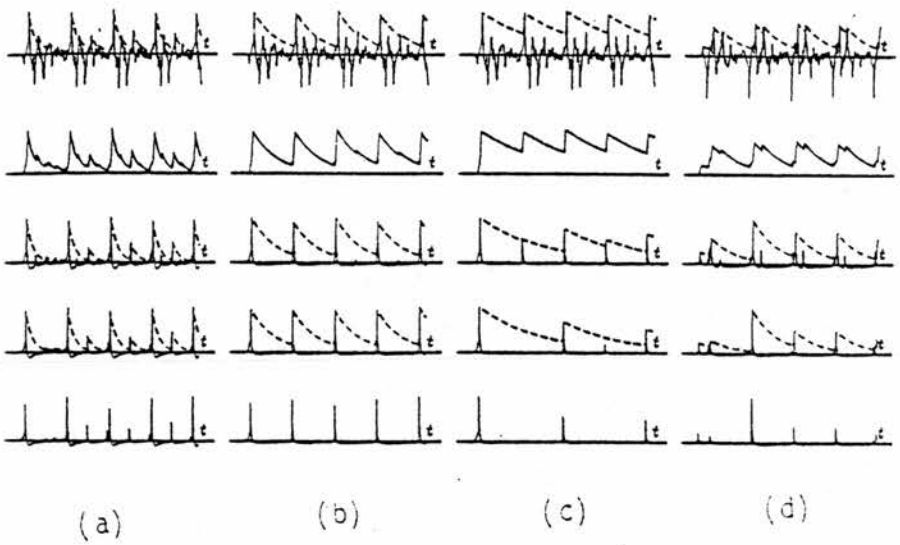


Figure 2.5 An example of time domain pitch detection based on an envelope modeling algorithm developed by Anderson (1960). For each part of Fig. 2.5 (a-d), there are 5 rows displaying the results of the envelope modeling process: Row 1 — original speech signal with dashed lines representing the initial application of the peak detection and exponential decay circuits; Row 2 — signal following initial application of envelope modeling circuit; Row 3 — signal following first application of differentiator circuit (the envelope modeling circuit is applied again); Row 4 — signal following second application of envelope and differentiator circuits; Row 5 — signal following third application of circuits. There are 4 parts to this figure: (a) — time constant for peak detector is too small; (b) — time constant is appropriate; (c) — time constant is too large; (d) — same time constant as in (b) but the signal has the wrong polarity. (Figure from Hess 1983:183).

waveform while suppressing secondary peak features. A differentiator circuit is then applied to the initial envelope model to further emphasize the principal peaks relative to any secondary peaks which may have been modeled as well (see, the third row Figure 2.5a-d). The application of the modeling and differentiation circuits can be done repeatedly until only the principal peaks remain from the original speech wave (the bottom three rows of Figure 2.5a-d represent 3 applications of the circuitry). Once the waveform has been simplified by the modeling process than the pitch periods can be measured from the interpeak distances. The period measurement is improved by a circuit which transforms the varying pulse shapes and amplitudes of the model into pulses of constant shape and size.

The most important factor for accurate measurement of periods by envelope modeling is the choice of time constant which controls the rate of decay of the basic extractor circuit. If the time constant is too small then secondary waveform peaks become prominent in the envelope model thus requiring additional processing of the signal by the circuitry. Figure 2.5a displays the results of envelope modeling with a time constant that is too small -- this constant leads to higher harmonic tracking by the device. A time constant which is too large leads to the suppression of principal peaks in the waveform since the exponential function intercepts the next peak well above the time axis (Dolansky 1955). Figure 2.5c shows that principal peaks have been lost due to the repetitive application of an exponential decay function with a large time constant -- this time constant leads to subharmonic tracking by the device. For signals with rapidly decreasing amplitudes, the time

constant of the decaying exponential model may appear too large. Hess (1983:188) notes that the solution for the proper choice of time constant is not easily found. In the first instance, the constant could be based on the bandwidth of a low first formant which is the primary waveform shaping the temporal structure of the speech signal. Since the overall amplitude of the signal can change drastically, the time constant would need to be small relative to the first formant bandwidth thus requiring repeated processing of the waveform.

It is clear from the above discussion that the envelope model of pitch periods suffers from both higher and subharmonic tracking of the waveform. This situation is due to the insufficient separation of the intraperiod temporal structure from the interperiod amplitude changes. In the case of the basic extractor, small constants lead to tracking of secondary peaks while large constants miss peaks. In the case of signal behavior, rising signal amplitudes lead to secondary peak detection while falling amplitudes lead to missed peaks.

Temporal Structural Analysis by Sequence of Extremes

The other type of temporal structure analysis includes pitch detectors which evaluate the Sequence of Extremes displayed in speech waveforms. These detectors directly examine the speech waveform for anchor points which reflect the periodicity of the voiced signals. Anchor points such as peak minima and maxima and zero-crossings are characteristics of the time domain waveform produced by the periodic excitation of the vocal tract by the vibration of the vocal folds. Sequence of Extremes pitch analyzers

apply heuristic models of the signal structure to the speech waveform; these models are constructed as computer programs which permit a variety of solutions to the time domain pitch detection problem. The main characteristic of these detection algorithms is data reduction at the basic extraction stage. The basic extractor produces a set of appropriate anchor points from which pitch period markers can be derived -- this process is characterized by iterative techniques of selection and elimination. The following is a sample algorithm for the Sequence of Extremes detector (Hess 1983:183):

> '1) Choose an appropriate feature and select all those samples of the signal to which this feature applies. Eliminate the rest of the samples (first step of reduction).
>
> 2) Select those of the remaining samples which are likely to represent a period delimiter, and reject those which are unlikely (second step of data reduction). Repeat this step if necessary.
>
> 3) Check all the selected delimiters as to whether they form a "regular" string of markers. Correct obvious errors.'

Due to an elaborate basic extractor, these systems will have a relatively simple pre-processor (usually a moderate low-pass filter to eliminate higher formants from the signal) or no pre-processor at all. As described in step 3 above, post-processing may include a correction routine to handle gross and fine marker detection errors. This type of algorithm requires a high degree of computer processing but operates fast and efficiently since most of the program consists of decision making and branching , and there is also the inherent data reduction in the basic extractor.

Hess has found it necessary to describe each one of the Sequence of Extremes algorithms in detail since each detector varies in its heuristic approach to the time domain signal. The following

sections are summaries of the algorithms classified as Sequence of Extremes analyzers. The reader is referred to Hess (1983) for a more comprehensive discussion as well as to the original authors. This area will also be discussed in more detail in the experimental section of this study in Chapter 3.

Simple Peak Detector and Global Correction — Reddy's (1967) algorithm is representative of those time domain detectors which process peaks in the speech waveform as anchor points for determining pitch period markers. The algorithm consists of two general stages: peak detection by the basic extractor followed by a global correction of pitch errors. The detector was designed for use in a speech recognition system and therefore was not required to work in an instantaneous mode. This design feature means that pitch periods are specified after analyzing the whole speech utterance.

Following the sample algorithm for the sequence of extremes extractor as set out above, Reddy's detector processes speech in the following manner (see Figure 2.6, after Hess 1983:202). The first step of data reduction is the retention of all local peak minima and maxima in a given segment of the speech signal (approximately 25 ms) and the elimination of the rest of the speech samples. Further data reduction is achieved by the determination of significant maxima and minima within the population of the local peaks. Several factors are involved in the assignment of significance to a given peak minimum or maximum. Firstly, the absolute maximum value is found amongst the local minima and maxima. Secondly, for a given signal polarity, a significant maximum or minimum is defined as follows: 1) the peak should be of the given polarity, 2) the peak should not be located within 2.5 ms of the previous significant peak of the

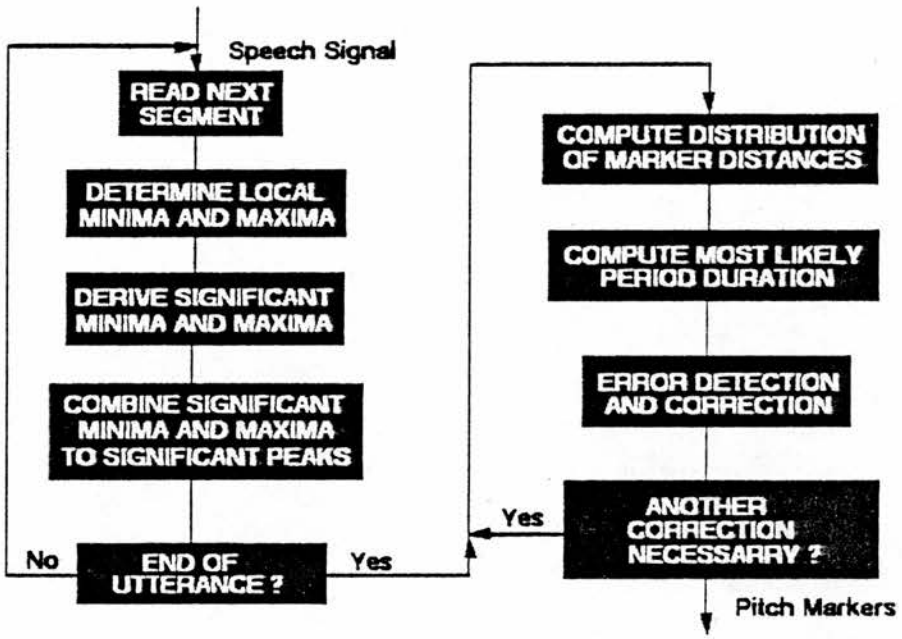


Figure 2.6 Block diagram of a time domain PDA by Reddy (1967) which processes peak minima and maxima in speech signals to determine period markers; this system includes global error correction of detected markers. (After Hess 1983:202).

same polarity, and 3) the peak should be (a) greater than .9 times the amplitude of the absolute maximum value or (b) greater than the linearly extrapolated value from the previous two significant peaks of the same polarity; if neither (a) or (b) condition is fulfilled within 13.5 ms of the previous significant peak then the significant peak is defined as the maximum of all the local peaks of the given polarity within that 13.5 ms segment of data. This second step of data reduction results in a set of significant peaks for both polarities of the speech signal. The final stage of data reduction is the elimination of extraneous significant maxima and minima to produce a set of significant peaks which delineate the pitch periods. Reddy noted that significant maximum peaks are usually followed closely by a significant minimum peak and that this feature would be a useful period marker. The close proximity of significant maxima and minima in the speech waveform reflects the vocal tract response to each glottal impulse. By definition, a significant peak is a significant maximum peak which has a corresponding significant minimum peak within 3.5 ms of its occurrence. Thus, the basic extractor analyzes a speech waveform which has not been pre-processed to produce a series of pitch markers based on peak minima and maxima.

The basic extraction of significant peaks in the speech waveform is followed by a global error correction of the markers using the technique of list correction. Reddy noted the necessity of marker correction since numerous incorrect marker placements were observed in comparison to a visual examination of speech data for pitch periods based on peak information. The marker placement errors found by Reddy are typical of TSA extractors and include 1)

holes -- markers which have been left out, 2) hops -- markers which have been misplaced and 3) chirps -- extra markers added where they should not be (Gold 1962; Reddy 1967; Hess 1983). Marker correction is often applied in the form of list correction of the various errors. List correction applies heuristics to the markers in a global fashion, either over an entire input utterance or at least a whole segment of continuously voiced speech. Hess (1983) outlined the general procedures for list correction of period markers. At the starting point, an initial estimate of the expected pitch period is required by the correction algorithm. In the case of Reddy's correction routine, the modal value of all the pitch periods of an entire utterance is used as the "most likely" pitch period length. The corrections are then completed in several steps (Hess 1983:286):

> '1) Check all the markers, accept those found regular, and mark the others for correction.
>
> 2) Scan all the markers again. Leave the regular ones unchanged. Correct chirps, hops, and holes where they can be corrected, and label the corrected markers regular where they fulfill the condition of regularity with respect to the expected period length. Repeat this step until there is no more change in the list. Label those markers irregular which do not fit into the correction scheme.'

The reader should refer to Reddy (1967) for the particular details of the global procedures applied in the peak detection system.

Hess (1983) noted several problems for the list correction of period markers in general, several of which apply to Reddy's algorithm in particular. The first problem is one of stability of the correction routine when numerous errors are present in the sequence of markers and a number of scans of the data are required to complete the corrections. Instability often arises from

contradictory requirements of the correction routine such as 1) correction of markers labeled as "regular" in a previous scan, 2) reinsertion of markers which were removed as unwanted chirps and 3) removal of markers inserted to fill holes in the marker sequence. The second problem has to do with the substantial delay in pitch extraction produced by the global correction procedure. Reddy's routine requires the analysis of the entire utterance by the basic extractor before the commencement of the global correction procedure. The necessity of a global correction routine of this complexity suggests that the algorithm by Reddy is too simple even for its restricted application. However, the program's simplicity and use of computer techniques demonstrate that the principle of structural analysis in the time domain does work for speech.

Zero-Crossings and Excursion Cycles -- Miller's (1975) algorithm is a useful example of structural analysis detectors which use zero-crossings as anchor points for determining the period durations in voiced speech (see Figure 2.7, after Hess 1983:203). In this system, data reduction procedures are based on the analysis of a temporal structural feature called the excursion cycle (EC). The excursion cycle is defined as all speech samples which occur between two consecutive crossings of the zero axis. A crude energy measure is derived for each EC present in the input utterance and this measure becomes the basis of a data reduction procedure which produces a signal at least one order of magnitude smaller than the original speech signal. The reduction routine is followed by a basic extraction and global correction of pitch markers similar to other TSA detectors.

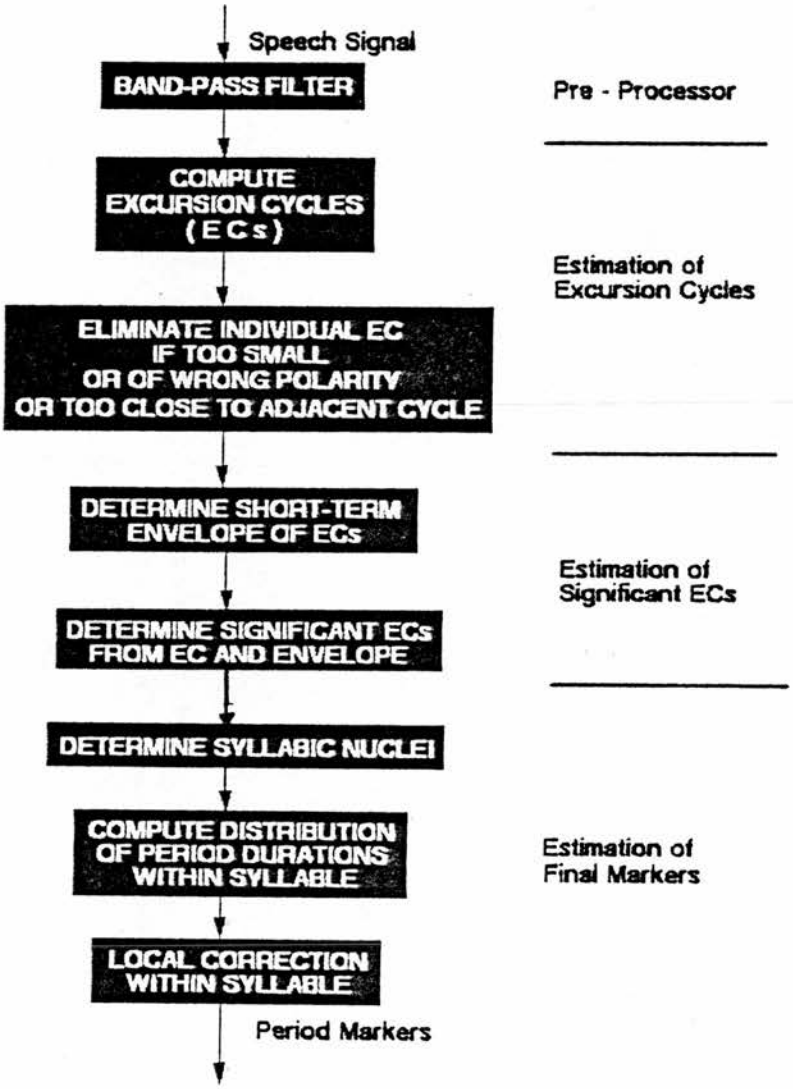


Figure 2.7 Block diagram of a time domain PDA by Miller (1975) which processes zero-crossings in speech samples to determine period markers. The analysis is based on a structural feature called the excursion cycle which is defined as all speech samples that occur between two consecutive crossings of the zero axis by the input signal. This system includes an element of global correction of the detected period markers. (After Hess 1983:203).

The data reduction routine of Miller's detector views the speech signal as consisting of a particular data structure. For voiced speech, the pitch period consists of a small set of ECs with the first EC being considered the significant feature for anchoring pitch markers. By definition, the beginning of each period is the first non-zero value of the significant EC. The task of the data reduction procedure is to limit the entire input signal's data structure to the significant ECs of the voiced speech segment. The data reduction procedure is as follows. Firstly, an energy measure is produced for each excursion cycle delineated by consecutive zero-crossings. Miller characterized the pitch period as having a large amplitude at the beginning of the period relative to the ending of the period (i.e. in the region of the significant EC). For voiced speech, each EC consists of a large amplitude and long duration compared to the ECs evidenced in unvoiced speech. The energy measure is calculated as the sum of all the samples within the excursion cycle. An energy threshold is applied to the ECs of both signal polarities which eliminates the segments of speech with low EC energy. Therefore, the first step of data reduction also serves as a voice/voiceless classifier. Secondly, Miller noted that the signal polarity of the significant ECs remain relatively stable throughout an utterance, thus the data is further reduced by excluding all ECs which are of opposite polarity to the significant excursions. The third stage of data reduction is based on the limitation of acceptable F0 periods to less than 500 Hz. For a given EC, if it has two neighboring excursion cycles of greater energy within 2 ms then the EC under examination is eliminated from the data structure. The results of the reduction methods is a series of significant ECs, each EC characterized by its starting

address, amplitude of the peak of the EC and the address of the peak.

The basic extractor detects pitch period markers from the significant ECs in a manner similar to other TSA devices. One unusual feature of the basic extractor is that markers are detected from syllabic units rather than straight sequential analysis. The extractor begins in regions of high energy of the ECs, usually in the vowel portions of syllables. Syllabic boundaries are then determined from the significant ECs followed by extraction of the individual markers. The basic extraction method plus a global correction of pitch errors prevents the system from operating in an instantaneous mode similar to Reddy's algorithm. The global correction procedure applies a list correction to whole voiced segments within the entire input utterance.

Hess (1983:207) noted that the Miller system is sensitive to second harmonic tracking due presumably to the high energy often found for the second harmonic in voiced speech. In addition, the zero-crossing just prior to the first EC in the period may be an unreliable period feature due to the phase relationship between the excitation signal and the output speech which is time-variant and unpredictable.

Mixed Feature Algorithms — Pitch detection algorithms included in this subcategory combine the freedom of TSA with features adopted from other areas. Mixed feature algorithms are considered to be the most convenient and powerful TSA detectors without being the most complicated. Hess (1983) labeled two algorithms in the literature as "mixed feature" including the parallel processing method

developed by Gold and Rabiner (1969) as well as the Tucker and Bates (1978) system which uses a non-linear transformation of the speech signal in the pre-processor. A brief description of the mixed features of the Tucker and Bates detector is given here while the Gold and Rabiner system is discussed in depth in the experimental section of this study in Chapter 3.

The algorithm developed by Tucker and Bates (1978) is similar to other TSA systems in that peak minima and maxima are the anchor points used to determine period markers within the time domain speech waveform. At the pre-processing stage, the algorithm applies a non-linear transformation to the speech signal in order to simplify the peak detection and elimination process of the basic extractor. The transformation is in the form of an adaptive center-clipper which emphasizes the principal peaks within the speech segment and eliminates the minor minima and maxima. The basic extractor then examines the remaining speech data for pulses which delineate the pitch periods. The location of each pulse is based on the peak value of each maximum and minimum remaining after the non-linear transformation. As each pulse is located, a pulse feature vector is derived which describes five properties about that pulse. Three of these features are primary features including the pulse amplitude (i.e. the peak amplitude), the pulse width (i.e. the duration between the two crossings of the center-clipper non-zero threshold) and the pulse energy (i.e. the sum of the square values within the pulse width). The two other pulse vector features are secondary properties including the pulse polarity and the pulse shape (defined as the ratio of the pulse amplitude to the square root of the pulse energy). The secondary features are not

affected by period-to-period variations of amplitude ~~which is~~ as found for the primary features. The similarity between pulse feature vectors is examined to determine those pulses which are likely delineators of pitch periods. The similarity between two given pulses for a given feature is defined as the absolute ratio of the differences between the two features to the sum of the two feature values

$$s_m = \left| (x_m - x_{m-v}) / (x_m + x_{m-v}) \right|$$

where X is a feature value for pulses m and m-v. A feature which is very similar for two pulses approaches a zero value while a feature which is very dissimilar for two pulses approaches a value of one. A vector of decision thresholds is then applied to the vector of similarity measures derived from the pulse feature vectors to eliminate unlikely pulses. The pitch periods are then determined from the remaining pulses. This algorithm does not require a global correction step following the basic extractor. Tucker and Bates (1978:600) noted that the mixed feature algorithm was applicable to a fundamental frequency range of 40 to 2400 Hz.

SECTION 2.1.3 — TIME DOMAIN PDAs WHICH PROCESS A TIME DOMAIN SIGNAL DERIVED FROM THE ORIGINAL SIGNAL

The two previous types of time domain pitch extractor have certain restrictions associated with their proper operation on speech signals. Fundamental harmonic detectors require the presence of the first harmonic in the signal and this waveform must be enhanced by linear and/or non-linear pre-processing techniques. On

the other hand, temporal structural analyzers must cope with a variety of signal structures which may require a manual pre-set for the expected F0 range (as found in analog versions) or elaborate post-processing to correct a variety of pitch extraction errors. These restrictions suggest an alternative solution to pitch period extraction in the time domain in which certain aspects of fundamental harmonic and TSA extractors are combined. The speech signal is transformed by an elaborate pre-processor and the resultant waveform is searched for pitch period markers by an elaborate basic extractor of the TSA type. The elaborate pre-processing produces a signal with significantly fewer temporal structures which simplifies the basic extraction task. Most intermediate devices are oriented to two features of periodicity in the speech signal. One type of device reconstructs the temporal structure of the input excitation signal by the use of an inverse filter to remove the vocal tract resonance characteristics from the speech waveform. The other type of intermediate device is termed an event detector. The event to be detected is the discontinuity in the speech signal associated with abrupt glottal closure during phonation. Various filtering techniques are applied to the speech signal to enhance the higher frequency signals in the waveform which have been simultaneously produced at the moment of glottal closure.

Simplification by Inverse Filtering

One intermediate time domain extraction method uses an inverse filter as a signal pre-processor to simplify the speech waveform prior to the basic extraction of pitch period markers. The basic principle of inverse filtering is that the excitation signal can be

reconstructed from the speech signal by a filter whose transfer function is the reciprocal of the vocal tract transfer function used to produce the original speech signal. The resultant signal is a simplified waveform which has to some degree preserved the excitation impulses used to produce the original speech waveform. For vowels and non-nasal sounds, the transfer function of the vocal tract can be modeled as an all-pole recursive filter -- the reciprocal of this transfer function is a non-recursive all-zero filter. Hess (1983:222) suggests a number of methods for designing and realizing an inverse filter for pitch period extraction. Firstly, a given segment of voiced speech can be analyzed for all the formants of the transfer function. These formants are then used to construct an appropriate system for filtering out the vocal tract response from the speech waveform. However, this system would be difficult to implement as an automatic analysis procedure. Secondly, a non-adaptive low-pass filter can be applied to a speech waveform to remove formant frequencies above the first formant. This filtered signal will then contain only one formant which needs to be located and modeled by an inverse filter.

The third possibility is the use of linear prediction analysis techniques to model the transfer function of a given sound and create an inverse filter from the transfer function. Linear prediction is a popular signal processing technique for speech analysis and synthesis which has been applied in a number of pitch detection systems described in other sections of this review. The reader is referred to the work of Makhoul (1975), Markel and Gray (1976), Rabiner and Schafer (1978) and Hess (1983) amongst others for complete discussions of the mathematical considerations and

applications of linear prediction. A very brief description of the technique as related to inverse filtering is given here. Linear prediction (Linear Predictive Coding; LPC) is a powerful parametric model which accounts for a number of speech signal characteristics in terms of a source/filter system. A series of statistical procedures are completed to model the transfer function of the system (i.e. the vocal tract resonances) as a recursive all-pole filter. This model filter is not a complete representation of the actual resonance and anti-resonance characteristics of the transfer function and therefore a certain amount of error occurs between the actual speech sample and its modeled equivalent. The error signal is often termed the residue. The best approximation of the actual signal is produced by the model with the least error upon output. In digital analysis, the LPC model produces a digital, recursive all-pole filter for the transfer function. Therefore, the LPC inverse filter is a non-recursive, all-zero filter which can be applied to the original signal from which it was derived to produce the residue signal (the residue signal should not be confused with original glottal waveform). LPC inverse filtering has been used for pitch detection in a couple of investigations of perturbatory behavior associated with pathological phonation. The study by Davis (1976), in which inverse filtering was used for pitch detection in the time domain, is discussed in detail in Sections 4.1.4 and 4.2.1 below.

The residue signal is a special waveform produced by a particular parametric representation of the speech signal. Due to the nature of the LPC model, the only safe assumption about the residue signal is that periodicity of the original signal is

maintained within it. As Hess (1983:226-227) points out, there are difficulties with this interpretation of the residue signal. Firstly, the LPC model is derived by a method which is meant to produce a recursive filter with minimal error characteristics. It is not specified in this method that the source impulse function of the speech signal is to be preserved in the residue. Therefore, it is unpredictable as to which properties of the speech signal is modeled by the filter and what properties are left to the residue. In the case of nasal sounds, for example, it is possible that the impulse function will be removed from the residue signal by the inverse filter. A further problem for inverse filtering techniques in general occurs when the first formant frequency coincides with the F0 as is sometimes the case for female speakers. Since much of the signal energy is present at this frequency region, it is sure to be modeled by an inverse filter technique with the result that the F0 has been filtered from the signal. This situation restricts the application of inverse filtering for pitch period extraction unless special logic is used to compensate for these problems.

Event Detection

According to the linear model of speech production, the voiced speech signal is the output of the vocal tract resonance system when stimulated by a periodic pulselike waveform. In this source/filter system, the higher frequency components (most notably in the formant regions) receive their main excitation at the instant of glottal closure (see, for example, Fant 1979a and b). The main task of event or epoch detection is to apply a temporal structure transformation to the speech signal in order to enhance those

waveform features which were evoked by the main excitation moment. The enhancement of the excitation features produces a simplified waveform from which the periodic events associated with glottal vibration can be detected by structural analysis or simple threshold basic extraction techniques.

In the first instance, the nature of the structural transformation for event detection is dependent on the location of the main excitation features — the excitation moment may occur at a zero-crossing or peak in the speech signal. This is a source of unreliability which is further compounded by any type of low-frequency phase distortion (e.g. in tape recorded data) which will make the main excitation point unpredictable for a given pitch period. Hess (1983:241) suggests a number of transformations for overcoming the phase problem in event detection. ~~The~~ This aspect can be removed from the event detection process by the application of a phase-splitting network or Hilbert transformation. One can confine the event detection to higher frequencies (e.g. above 1 KHz) which have not been overly phase-distorted or to narrowbands of frequencies in the formant regions (it is assumed that these formant locations must be determined prior to the event analysis).

One example of an event detector for pitch detection was created by Smith (1954, 1957 as reported in McKinney 1965:120) which uses a filter bank to analyze the speech signal. The speech signal is input to a filter bank consisting of 32 second order bandpass filters spread across a given frequency range. The filtered output from each bandpass filter is full-wave rectified resulting in a demodulated version of the waveform. All the bandpassed, rectified signals are then summated to produce a waveform with the pitch

periodicity as its chief temporal characteristic. This simplified signal is the input to a basic extractor (in this example, a temporal structural analyzer) which detects the instants of main excitation. The temporal structure of the summated waveform is due to two factors. Firstly, all pitch-related components in the original speech signal are phase-coherent and therefore add together to reflect the rise and decay of energy with each impulse. The second factor is that the remaining non-pitch-related components are phase-incoherent and cancel each other out upon summation of the rectified signals. These points can be further demonstrated from a similar event detector implemented by Yaggi (1962; 1963 as presented in Hess 1983:232-233). Following the rectification stage, each output from a bandpass filter is smoothed by a low-pass filter prior to the final summation (see Figure 2.8, from Hess 1983:233). The low-pass filter cutoff is presumably set to a cutoff frequency just above the upper end of the fundamental frequency range of interest. The consequence of filtering, rectification and smoothing is an amplitude demodulation of each bandpass filtered signal. This effect is particularly notable if one of the bandpass channels contains a formant frequency since the waveform will contain a considerable amount of energy associated with the main excitation. The demodulation of a waveform closely related to a formant frequency produces its temporal envelope which follows the periodicity of the original excitation signal. The simultaneous excitation of all the formants by the main excitation will be evidenced in the coherent summation of the temporal signals associated with the vocal tract resonances.

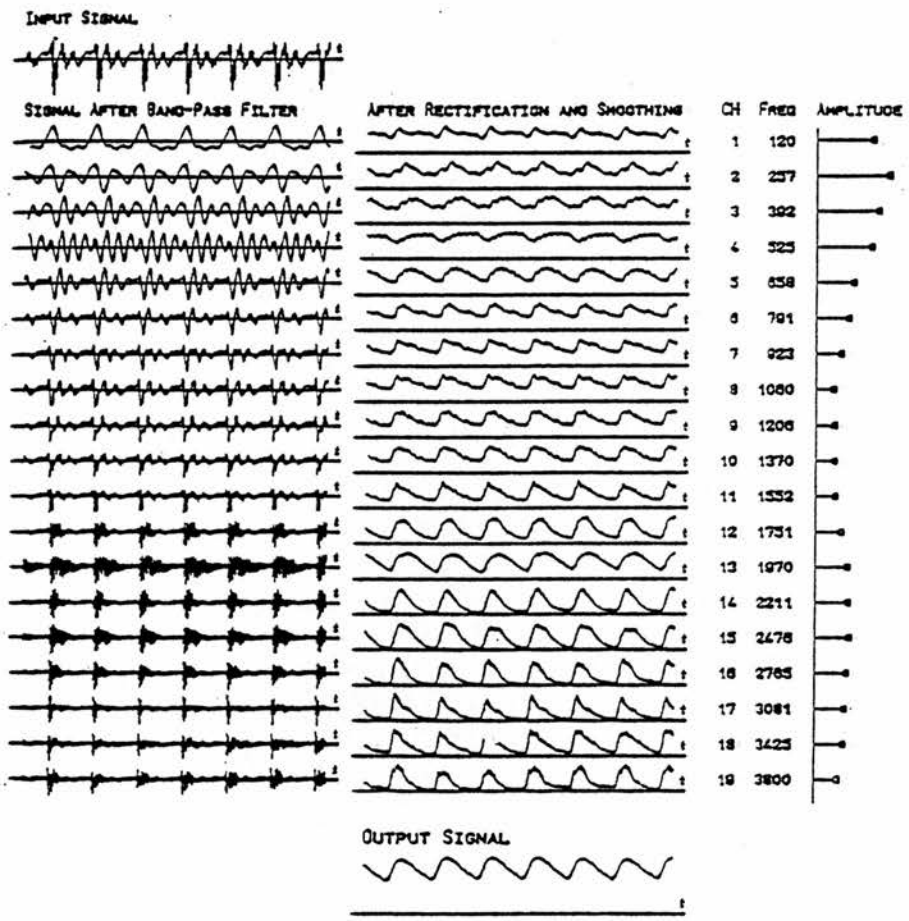


Figure 2.8 An example of time domain pitch detection based on the event detection algorithm of Yaggi (1962). The input signal is passed through a 19 channel bandpass filter bank (CH -- channel number; FREQ -- center frequency of channel in Hz); the output of each channel is full-wave rectified and smoothed; the outputs from the filters are summated to produce a signal which displays the periodicity of the original input signal. The amplitude of the signal in each bandpass filter is shown on a linear scale. (Figure from Hess 1983:233).

The interaction of the F0 and the first formant for certain voices (mainly high-pitched females and children) is not as extreme a problem for event detection as for inverse filtering. For the event detector, filtering and rectification stages produce a pure DC signal with periodic increases in energy at half the period of the F0-F1 coincidence frequency. If enough data is available from the unaffected higher formants then the F0-F1 signal will only have a slight detrimental effect upon the overall summated signal. The problems associated with F0-F1 interaction as well as secondary excitation or weak discontinuities in the glottal waveform can be avoided by restricting the event detection to frequencies in the higher formant regions. The damping of the higher formants is in general broader than the response of the low-frequency formants and therefore should respond to weakened instances of glottal excitation.

MULTICHANNEL SOLUTIONS TO PITCH DETECTION IN THE TIME DOMAIN

A number of short comings have been noted for most of the time domain PDAs which use a single channel for determining pitch markers in voiced speech. The ability to process speech waveforms in parallel through several channels should improve the performance of these detectors by taking advantage of the redundant information present in these signals. Multiple processing components may exist in any or all of the basic components of the time domain PDA. According to Hess (1983:242), multichannel PDAs may be subclassified along the following three principles:

> 1) The principle of main and auxiliary channels -- This PDA consists of two channels including an accurate main channel which requires tuning to a proper frequency range and a crude auxiliary channel which does not require

tuning and directs the main channel to the correct frequency range, For example, certain fundamental harmonic extractors follow this principle by using a tunable filter which is controlled by an open-loop circuit.

2) The subrange principle -- A PDA following this principle consists of a number of similar or identical PDAs, each one operating over a subrange of the overall fundamental frequency range of interest.

3) The multiprinciple PDA -- This is a PDA consisting of a number of channels with each one operating independently of the others. Each channel may process a different parameter or the same parameter but with the application of a different set of criteria. Hess (1983) includes the event detectors by Smith (1954) and Yaggi (1962) under this principle; it is assumed that these PDAs are classified in this way since the extraction of a singular given event (i.e. the moment of glottal impulse) is completed by a number of parallel pre-processing channels. It can be argued that PDAs which use filterbanks in the pre-processor also satisfy the subrange principle.

The parallel components of the multichannel PDA may exist in any or all sections of the time domain extractor.

The use of multiple channels for pitch extraction does introduce problems of its own. Firstly, the selection of the correct channel becomes problematic when multiple channels are active for a given sample of speech (e.g. the presence of higher harmonics in the other channels). Two general approaches to the channel selection problem have been used in the multichannel PDA. For PDAs following the subrange principle, a minimum-frequency selection principle is often applied such that the lowest active channel (i.e. the lowest frequency) is selected and all other channels are blocked. In the case of multiprinciple PDAs, one observes the error behavior displayed within individual channels (usually at the basic extraction stage) and then creates a selection routine according to the observed behavior. The second problem is that of marker synchronization in which the phase relationships between the markers present in the various channels may be lost.

This is not a problem for all multichannel PDAs but when it does occur, it can be very difficult to overcome.

PDAs following the Subrange Principle

As discussed in the section on fundamental harmonic extraction, one solution to the limited F0 range of operation for these detectors is the use of a tunable filter in the pre-processor tuned either by manual or automatic controls. In the multichannel approach, the tunable filter is replaced with a number of fixed pre-processor filters, each filter covering a subrange of the total frequency range of interest. One example of this technique was developed by Kubzdela (1976). One notable problem associated with this type of channel selection is the introduction of phase distortion into the resultant time domain waveforms.

Multichannel PDAs operating with the subrange principle adopt two methods of channel selection. The first method is the minimum-frequency selection routine which was discussed above. The other method is optimum frequency matching which selects a channel according to a given logical condition. The minimum-frequency selection principle is the more favored routine since it should be insensitive to higher-harmonic tracking and does not cause significant delays in the processing of data. This method may be sensitive to low-frequency spurious signals and rapid transitions with strong low-frequency components.

An example of the multiprinciple approach to time domain pitch detection, created by Gold and Rabiner (1969), will be discussed in detail in the experimental section of this study in Chapter 3.

SECTION 2.1.4 — SUMMARY -- TIME DOMAIN PDAs

The preceding sections were a brief review of time domain pitch detection algorithms. A time domain PDA is one in which the basic extractor of the system extracts period markers from a signal with a time-base equivalent to that of the original input speech waveform. The time domain PDA must be capable of processing a variety of signal structures. Periodicity is evidenced in a number of ways within the speech waveform and time domain PDAs have been designed to take advantage of these characteristics. One mode of operation in the time domain is the extraction of the fundamental harmonic component evidenced in the waveform for voiced speech. The fundamental harmonic is highlighted by a variety of non-linear and/or linear techniques in order to improve its detection by thresholding procedures. Another mode of operation in the time domain is the analysis of the overall temporal structure of the speech waveform for the presence of repetitive signal characteristics. Anchor points such as peaks and zero crossings are used as markers of structural repetition within the speech waveform. Other time domain PDAs take advantage of the two previously mentioned modes of operation by applying temporal structural analysis techniques to waveforms which have been produced by elaborate pre-processing methods. In this case, the pre-processed signal shows properties which may be similar to the input excitation signal used to stimulate the vocal tract in the production of voiced speech sounds. A number of multichannel solutions have been used to improve pitch detection by the use of redundant periodic information within the temporal speech signal.

The main advantage of pitch detection in the time domain is the ability to determine individual pitch periods within the speech waveform. These individual periods may be captured as they change rapidly in duration or if they are slightly irregular in their production. The main disadvantage of time domain analysis is that the speech waveform itself can be corrupted by noise and low-frequency phase distortion. The signal must also be properly resolved at the sampling stage to insure accurate results.

SECTION 2.2 -- SPECTRAL DOMAIN PITCH DETECTION ALGORITHMS

General Realization of Spectral Domain PDAs

The general realization of a spectral domain PDA is shown in the block diagram of Figure 2.9 (after Hess 1983:348). The detection process may begin with optional pre-processing of the input speech signal by linear or non-linear means -- this pre-processing simplifies the extraction task by limiting vocal tract resonance characteristics from the speech waveform. The speech waveform is then subdivided into short segments of suitable length for the short-term analysis of the parameters of interest for pitch detection. Each segment of speech is then transformed into the new domain which focuses the pitch information contained within the data. The basic extractor is usually a peak-detecting algorithm which locates the major peak -- the voiced/voiceless decision is usually completed at this point, based on the strength and location of the peak. The post-processor provides the pitch period or F0 parameters in a format appropriate to a given speech analysis task.

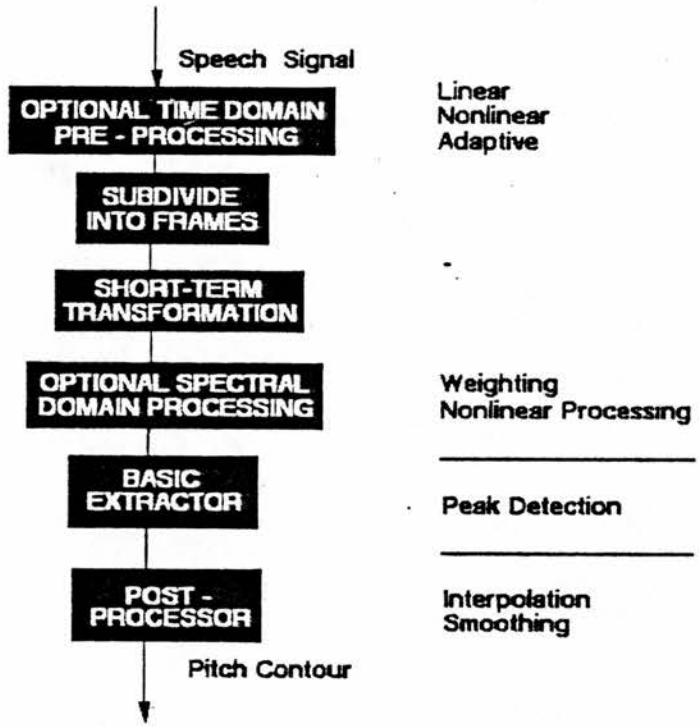


Figure 2.9 Block diagram of a typical spectral domain pitch detection algorithm. (After Hess 1983:348).

Spectral Transformation and the Focusing Effect

Pitch detection algorithms classified as spectral domain detectors operate in domains other than the original input signal's time domain. These other domains are reached by a short-term transformation of a selected segment of speech waveform at the pre-processing stage. All the useful PDAs of the spectral domain type demonstrate an ability to focus pitch information. This property can be likened to a concave mirror in which all indicators of periodicity within the short-term spectrum of a speech segment are concentrated into a single principal peak measure (maximum or minimum). The principal peak is detected by the basic extractor and interpreted as the average pitch period or F0 for the particular segment of speech.

Short-term Analysis Techniques

A number of considerations must be made when applying spectral domain algorithms to pitch detection of speech. In short-term analysis procedures, windowing techniques are used to determine local momentary parametric values from the larger time-variant speech signal. Here, a short segment of speech is selected by an appropriate window function, a set of parameters (the frame) is derived from the data within the windowed segment by an analysis procedure, and then the window is moved along the speech waveform in time by a specified frame interval (also known by its inverse -- the frame rate) for further short-term analysis. The use of a window function has several issues associated with it. Firstly, the data within the window, determined by the window length, is assumed to be quasi-stationary and representative of a given speech parameter at

that moment of analysis. For pitch period estimation, a balance must be met between these two requirements when choosing the time interval contained within the window -- 1) a minimum length of 2 pitch periods is required to produce accurate (i.e. representative) results using short-term analysis techniques and 2) a maximum window length should be chosen such that the data within remains stationary or quasi-stationary in comparison to the slow-moving parameters found for continuous speech. Hess (1983:345) suggests that the two requirements are just compatible for spectral domain pitch detection of segments 20 to 50 ms in length. Secondly, spectral distortion occurs when the window function is convolved with the short-term signal function. The spectral peaks of the original speech spectrum may be smeared by the principal bandwidth properties of the window and spurious peaks may be introduced into the spectrum. Various window functions (e.g. Hamming, Hanning and Kaiser windows) have been developed to limit the spectral distortions. Thirdly, the short-term analysis of speech requires that the data outside the windowed interval of interest be defined in relation to the data within the window. Two methods are typically used in pitch detection for defining these external data points. In stationary analysis, all samples outside the window are forced to zero resulting in a function which displays finite energy and duration. A long interval of several pitch periods is required for stationary analysis. A majority of spectral domain pitch algorithms operate in a stationary analysis mode. In non-stationary analysis, speech samples adjacent to the windowed segment of speech are used to derive parameters within the interval. This method is applied to time-variant signals and requires relatively short segments of speech within the window. Finally, the frame interval for shifting

the window should accurately follow the movements of the parameters derived from continuous speech. For pitch extraction, the bandwidth of the pitch parameters is relatively small compared to the bandwidth of the input sample speech. Hess (1983:345) suggests a frame update interval of 10 to 40 ms will accurately track most pitch movements using a spectral domain pitch detector.

Each short-term analysis of a windowed segment of speech produces an average measurement of pitch period or F0 for all pitch data contained within the section. For a repetitive analysis of a speech waveform, the output of the spectral domain detector is a sequence of average period estimates over time, the number of estimates being dependent on the frame rate.

Phase Insensitivity and Noise Resistance

One of the consequences of the short-term transformation procedure is a loss of the phase relationship between the original speech signal and the resultant pitch estimates extracted by a spectral domain PDA. Therefore, spectral domain pitch detectors are not capable of tracking period-to-period features of source excitation which are displayed in the voiced sections of time domain speech signals. However, this phase insensitivity does mean that spectral domain PDAs are not strongly affected by phase distortions of the original speech waveform. In addition, the focusing property of the short-term transformations make these PDAs resistant to noise and signal degradation (particularly in the case of bandlimited speech signals where the fundamental harmonic is not present).

The General Modes of Pitch Detection in the Spectral Domain

The various useful spectral transformations of the speech waveform which produce the desired focusing effect of the spectral pitch information have broadly determined the modes of operation for spectral domain PDAs. Based on these spectral transformations, spectral domain PDAs may be summarized from Hess (1983) as operating in the following manners:

1) Spectral domain PDAs which determine period data from a given correlation function. Correlation-based algorithms produce transformations which represent time as a delay (lag) between 2 correlated input signals. These detectors have been described as time domain (see, for example, Rabiner and Schafer 1978) since the resultant function produced by the short-term transformation has a time baseline. This ambiguity is resolved by noting that the transformation of the original signal is into a time domain unlike the original speech waveform processed by time domain devices. Two of the more useful spectral transformations are the autocorrelation and average magnitude difference functions. For the autocorrelation function, high correlations will occur at lags equal to the pitch period and multiples of that period duration. For the average magnitude difference function, a minimum difference will be found at a lag equal to the period. The task of the basic extractor is to determine the correlation lag equivalent to the pitch period.

2) Spectral domain PDAs which determine F0 or period data from frequency domain representations. Included in this section are frequency domain techniques which directly examine the spectrum produced by a short-term Fourier analysis of an input speech segment for indicators of periodicity -- these indicators may be in terms of frequency or delay-time. Frequency domain PDAs examine the harmonic structure of a spectrum for indicators of F0. Cepstral analysis uses a further frequency transformation of the frequency spectrum to focus the harmonic energy into one time-based peak.

The following sections discuss these general approaches to pitch extraction in the spectral domain.

SECTION 2.2.1 -- SPECTRAL DOMAIN PDAs WHICH USE CORRELATION FUNCTIONS

Spectral domain detectors that use a correlation for a transformation create a function which reflects the degree of correlation between two input signals. This function is time-based with the amount of delay as the independent variable and the degree of correlation as the dependent variable. The two types of correlation function described here are 1) autocorrelation which determines the degree of similarity between an input signal and a delayed version of itself and 2) anticorrelation which estimates the degree of dissimilarity between a signal and its delayed version. These two correlation methods are closely related since high similarity, that is, low dissimilarity is displayed by a quasi-periodic speech signal.

Autocorrelation PDAs

The degree of similarity or agreement between two input functions can be measured as a correlation. A special case of correlation is the degree of agreement between the signal and itself, that is, the autocorrelation function (ACF). A delay (often called the lag) between 2 input channels of the correlator is the independent variable of the autocorrelation function -- the amount of correlation found for a given lag being the dependent variable. Periodic or quasi-periodic signals evidence great similarities in waveform and high correlation coefficients are found for lags equal to one or more multiples of the signal's pitch period. Therefore, the basic extractor of the elementary ACF pitch detector would use a peak-picking strategy to determine the lag with a high correlation which represents the pitch period.

The transform of the time domain signal by the ACF results in a new time-dependent function which exists in the lag domain. For a discrete time signal, the ACF is defined as (Rabiner and Schafer 1978:141)

$$AC(k) = \sum_{-\infty}^{\infty} X(m) \cdot X(m+k)$$

In this function, the autocorrelation coefficient AC is the summation of the input signal x multiplied by a delayed version of itself where k represents the lag. It should be noted that this formula has not been completely defined for a window-dependent short-term transform of a signal. It is assumed for the ACF that all values outside the short-term window are forced to a value of zero. The following properties of the ACF which contains the speech signal energy as a special case are summarized from Rabiner and Schafer (1978:141) as follows:

1) If the original signal is periodic, then the ACF of that signal is also periodic with the same period, i.e. AC(k) = AC(k+P) where P is equal to the pitch period.

2) The ACF is an even function, i.e. AC(k) = AC(-k).

3) The maximum value of the ACF is found for k = 0.

4) The quantity AC(0) is equal to the average power of random or periodic signals.

The first 3 properties suggest that the ACF attains maximum correlations when the lag is equal to zero, the pitch period, and multiples of the pitch period. Therefore, the first maximum of the ACF beyond the zeroth lag can be used to determine the duration of the pitch period without regard to the original time base of the signal.

Figure 2.10 (from Rabiner and Schafer 1978:143) displays ACFs for 2 voiced segments (2.10a and b) and an unvoiced segment (2.10c). A window length of 401 samples was used to segment a section of speech sampled at 10 KHz, the ACF was calculated for 0 to 250 lags (the ordinate). Note that the ACF values have been normalized to the energy at the zeroth lag (i.e. the average power of the segment). In the 2 voiced examples, peak correlation values are displayed at lags equivalent to zero and multiples of the pitch period -- the first peak beyond the zero lag in 2.10a is at lag equal to 72 (7.2 ms/138.9 Hz) and in 2.10b at lag 58 (5.8 ms/172.4 Hz). It can be seen that the ACF tends to slope off in energy as the lag number increases since fewer and fewer speech samples are included in the correlation. For the unvoiced segment in 2.10c, no clear indication of periodicity is present which is the expected result for the aperiodic structure of the original waveform.

In the so-called "ordinary" (Rabiner 1977; Hess 1983) ACF pitch detector, the extractor functions in the following manner. Following some (optional) moderate low-pass filtering, the pre-processor completes a short-term transform of a frame of speech from the time to the lag domain. The basic extractor is a peak detector used to locate the first non-zero lag maximum in the ACF, this peak reflecting the focused pitch information caused by the transform. The ACF provides the peak detector with a useful reference point for determining a threshold, that is, the maximum energy as measured at lag equal to zero. If a peak is detected by a pre-determined threshold, the segment is labeled voiced and the period determined form the lag location of the maximum; otherwise the segment is classified as voiceless. Rabiner (1977) noted that

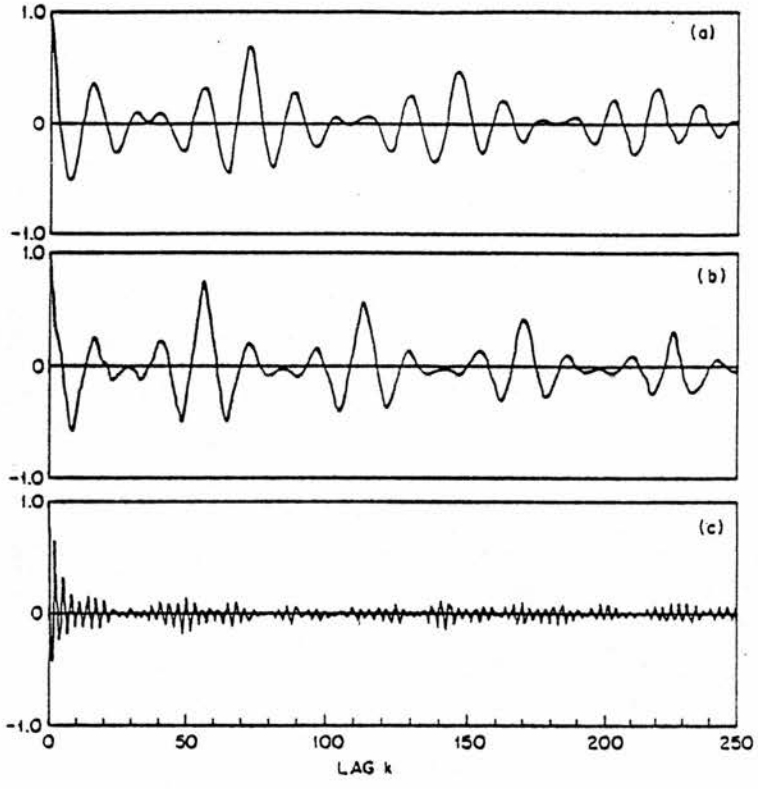


Figure 2.10 Three examples of autocorrelation functions computed for frames of voiced speech (a) and (b) and unvoiced speech (c). The lags range from 0 to 250 (ordinate) and were obtained from intervals of 401 samples of speech; each function has been normalized by the zeroth lag. For the voiced examples (a) and (b), peaks occur in the functions at multiples of the input signal's pitch period. (Figure from Rabiner and Schafer 1978:143).

the ACF is useful for pitch detection since it can be applied directly to the speech waveform in a computationally straightforward manner and is not sensitive to any phase distortion of the signal.

The relationship of the ACF to the power spectrum of the signal is suggested by the phase insensitivity of the transform. That is, the ACF is related to the power spectrum via an inverse Fourier transform. This property reveals the major drawback to the use of the ordinary ACF pitch detector -- this is the tendency of the detector to track harmonics or sub-harmonics in the signal since the ACF also displays the energy of the resonant structure present in the original speech (Hess 1983). As can be seen in Figure 2.10a and b, there are many peaks in the ACF, most of which are due to the damped oscillations of the vocal tract response which shapes each pitch period of the time domain speech waveform. Rabiner and Schafer (1978:150) noted that in some instances (e.g. rapidly changing formant frequencies) non-pitch period peaks are greater in correlation energy than values associated with the period. Therefore, most useful ACF extractors rely on some form of pre-processing of the time domain signal to limit the effects of the formant structure on the autocorrelation.

Sondhi (1968) found that non-linear and linear pre-processing of the speech waveform, which spectrally flattens the signal, greatly improved the performance of the ACF detector. Spectral flattening removes the vocal tract influences in the waveform and brings each harmonic to the same amplitude level, similar to the harmonic structure of a periodic impulse sequence. The autocorrelation of the spectrally-flattened signal shows distinct peaks at the lags equivalent to multiples of the pitch period (and

the zeroth lag) while the remaining energy is greatly reduced. Sondhi investigated three types of spectral flattening including band-pass filtering, band-pass filtering plus minimum phase compensation (two forms of adaptive linear pre-processing) and center-clipping (a form of non-linear pre-processing). An an example of the usefulness of pre-processing to autocorrelation pitch detection, the non-linear center-clipper will be discussed here. Figure 2.11 (from Sondhi 1968:265) displays an example of a center-clipper applied to a segment of speech. In the upper part of the figure, two clipping levels +/- ka0 are applied to the speech waveform (k is a pre-determined percentage of the peak amplitude a0). The energy exceeding the clipping levels (the shaded regions) are kept and compressed to the original time baseline as displayed in the lower portion of Figure 2.11. The resultant ACFs for a number of center-clipped waveforms are displayed in Figure 2.12 (from Sondhi 1968:265) for several segments of continuous speech (30 ms window lengths, 15 ms frame intervals, lags up to 15 ms) which were Hamming windowed prior to autocorrelation analysis. It can be seen that center-clipping is a simple and efficient technique for removing the effects of formants from the ACF thus improving pitch period peak detection. Sondhi found that autocorrelation pitch detection combined with spectral flattening worked well in the presence of high-pass filtering and broadband noise. Listeners rated synthetic speech which used ACF extracted pitch periods as the source parameter as good in quality. The problems of period doubling or voiced-to-unvoiced errors were not found for this autocorrelation detector which used spectral flattening in the pre-processor.

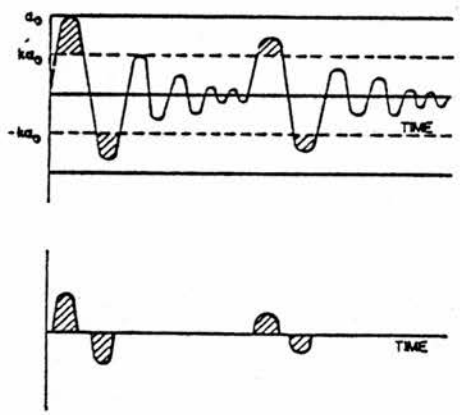


Figure 2.11 An example of non-linear pre-processing of a speech signal using center-clipping and compression. (Figure from Sondhi 1968:265).

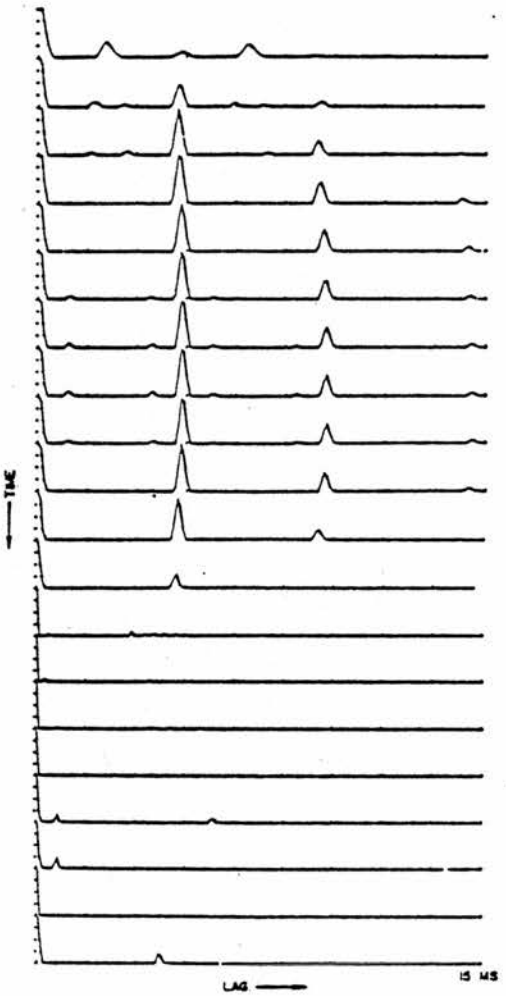


Figure 2.12 Examples of autocorrelation functions derived from speech signals which have been center-clipped prior to the short-term transformation. Note that much of the effects of the formants have been removed from the autocorrelation functions. (Figure from Sondhi 1968:265).

Most of the successful ACF detectors use some form of pre-processing to improve their performance. For example, Dubnowski, Schafer and Rabiner (1976) built a hardware ACF detector which includes an energy adaptive center-clipper/infinite peak clipper pre-processor which greatly simplifies the computation of the ACF. A versatile ACF detector with non-linear pre-processing was developed by Rabiner (1977) for use with a variety of input speech signals. In this detector, 3 different non-linear processors (center-clipper, center-clipper plus compression and center-clipper/infinite peak clipper) may be paired in a number of combinations to process the two input channels to the correlator. A well-known ACF detector created by Markel (1972) uses an adaptive linear pre-processor to spectrally flatten the input signal. The simplified inverse filter transform (SIFT) uses an LPC derived inverse filter to limit the influence of the vocal tract response from the speech signal. The resultant residue signal from the inverse filtering (in some ways equivalent to the source input to the vocal tract) is then autocorrelated for pitch detection.

Anticorrelation PDAs -- The Average Magnitude Difference Function

The autocorrelation function is useful in pitch detection since a periodic signal will evidence high correlations in the lag domain at delays equal to multiples of the pitch period. Another function termed "anticorrelation" by Hess (1983) quantifies the differences between 2 input functions. That is, two functions which display great similarities will not exhibit great differences. Thus, a difference function also measured in the lag domain will evidence minima (i.e. small differences) at delays equal to a pitch period

or multiples thereof for periodic signals. The basic extractor applied to a difference function would require a minimum peak-picking routine to detect the lag with smallest difference which is located at the pitch period.

The average magnitude difference function (AMDF) is a special case of generalized distance functions which has been used for pitch extraction. The AMDF is defined by Hess (1983:373) as

$$\mathrm{AMDF}(d) = \frac{1}{K} \sum_{n=q}^{q+K-1} \left| X(n) - X(n+d) \right|$$

In this function, the AMDF coefficient is the average of the magnitude difference of the input signal x with a delayed version of itself where d represents the lag. This formulation has been defined for a k length window which has been shifted to a starting sample with address n. Hess (1983) notes that the AMDF is a non-stationary function since the number of samples involved in the correlation is larger than the window length and therefore the signal is not assumed to be zero outside the window. A strong minimum is expected in the AMDF where the lag is equal to a multiple of the pitch period -- this minimum is equal to zero for a signal which is exactly periodic. The AMDF is phase insensitive since the harmonic structure is removed without regard to the original time origin of the speech; in a sense, the AMDF acts like a comb filter with minimum energy difference as a measure of best fit. The AMDF has been found to be sensitive to intensity variations, noise and low-frequency spurious signals. Computational advantages of the AMDF are the lack of required multiplications, scaling and double precision to produce the function (Rabiner and Schafer 1978) as well

as the relatively small window length associated with non-stationary processing (Hess 1983).

Figure 2.13 (from Rabiner and Schafer 1978:150) displays AMDFs for three speech segments (2 voiced and 1 unvoiced examples) which were also presented in ACF form in Figure 2.10. Each AMDF was calculated for lags of 0 to 250 and each resultant function has been normalized to 1.0. The voiced segments in Figure 2.13a and b evidence strong minimum energy differences at lags equivalent to multiples of the pitch period, the locations being similar to the peak correlation delays of the associated ACFs. In the unvoiced segment, no clear minimum difference which would indicate periodicity is seen in Figure 2.13c.

A simple version of the AMDF pitch extractor would use the following scheme. The periodic information of a short segment of speech is focused by a short-term transform using the AMDF. The basic extractor is a threshold function which determines the minimum with the smallest difference in the AMDF. If a minimum surpasses the pre-determined threshold, than the segment is classified as voiced and the pitch period estimated from the location of the minimum, otherwise the segment is labeled as unvoiced. Due to the sensitivity of the AMDF to intensity variations and noise, there is no readily available reference which can be used for determining a threshold for voiced/voiceless classification.

The AMDF has been applied directly to the speech signal for pitch detection (see, for example, Moorer 1974; Ross, Shaffer, Cohen, Freudberg and Manley 1974). Linear and non-linear pre-processing has been applied to the speech waveform to spectrally

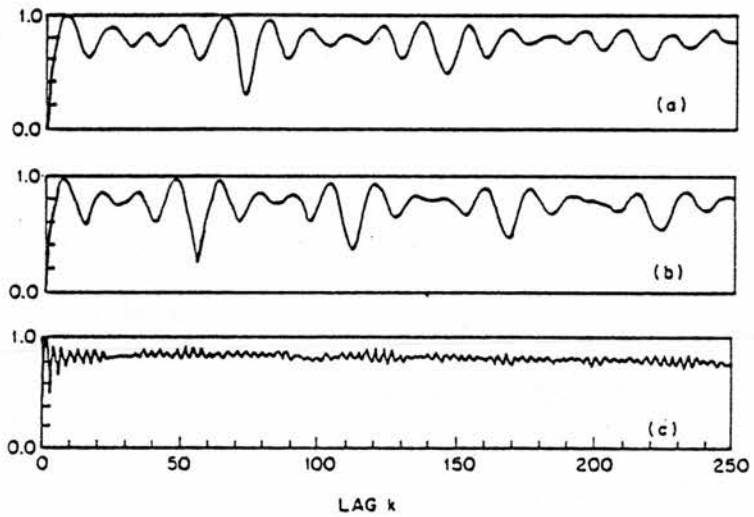


Figure 2.13 Three examples of average magnitude difference functions computed for frames of voiced speech (a) and (b) and unvoiced speech (c). These functions are derived from the same signals used in Fig. 2.10. The lags range from 0 to 250 (ordinate) and were obtained from intervals of 401 samples; each function has been normalized to a value of 1.0. For the voiced examples (a) and (b), minima occur in the functions at multiples of the input signal's pitch period. (Figure from Rabiner and Schafer 1978:150).

flatten the harmonic structure. Un and Yang (1977) applied the AMDF to spectrally flattened speech residue produced by LPC inverse filtering (adaptive linear processing).

SECTION 2.2.2 -- SPECTRAL DOMAIN PDAs BASED ON FREQUENCY DOMAIN REPRESENTATION OF THE SPEECH SIGNAL

A number of spectral domain PDAs which use a frequency domain representation of the speech signal for pitch detection are presented in the following sections. The pre-processor used to reach the frequency domain usually consists either as a series of bandpass filters spread over the frequency range of interest or in digital form as a discrete Fourier transform (the Fast Fourier transform being the computationally efficient technique). A special case of a frequency domain pitch detector is cepstral analysis which uses multiple spectral transformations to produce a time-based cepstrum from which pitch period data may be detected. The other frequency domain PDAs to be discussed are based on direct analysis of the harmonic structure of the frequency spectrum to determine fundamental frequency data.

Cepstral Analysis

Cepstral analysis is a frequency domain method for deriving speech parameters from the acoustic speech waveform. The power spectrum of a segment of voiced speech is treated as a linear product of source spectrum and the combined filtering effects of the vocal tract (Wakita 1976). In the time domain, the functions describing the source input and the vocal tract transfer properties

are said to be convolved together to produce the output speech signal. The separation of source and filter characteristics is called deconvolution. The deconvolution process can be greatly simplified by treating the speech signal as the result of an additive process rather than a multiplicative one. Transformation of the power spectrum to the log spectrum produces a function which is the sum of the source log spectrum plus the filter log spectrum. The transformation of periodic speech data into its log spectrum demonstrates "periodic" behavior in the frequency domain (Schroeder and Atal 1962). Figure 2.14a (from Noll 1967:296) displays the combined effects of voicing and filter resonances in a log spectrum of a segment of voiced speech. The "periodicity" of the log spectrum is seen as a regularly spaced "high-frequency ripple" in the frequency domain. The log spectrum's periodicity is characteristic of the quasi-periodic vibrations of the vocal folds during voicing. Noll (1967:295) noted that the "periods" of the log spectrum are the reciprocal of the period of the original time signal (i.e. the fundamental frequency). The log spectrum also contains the effects of the filter resonances upon the voicing information. This is seen as a "low-frequency ripple" which shapes the overall curve of the "high-frequency" ripple. It can be seen that the effects of the source and filter spectra may be distinguished by separating the high and low frequency ripple patterns. If the log spectral curve is viewed as just another waveform, then ripple separation becomes a matter of calculating a new spectrum which represents the behavior of the two different frequency ripples. The cepstral technique of pitch extraction requires that an inverse Fourier transform of the short-term log spectrum be completed. The resultant spectrum is called the

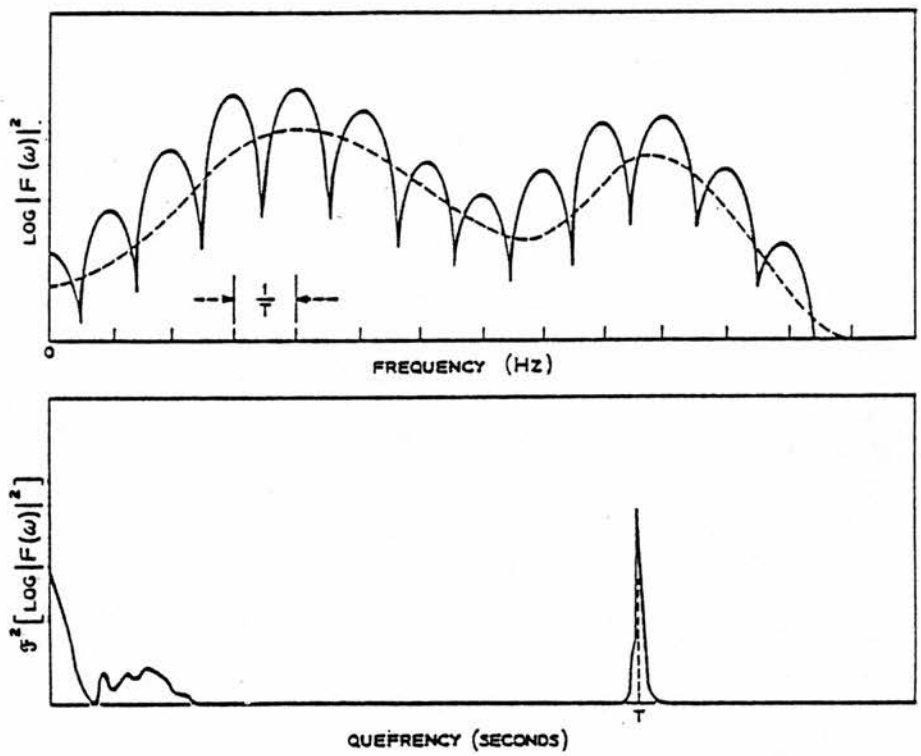


Figure 2.14 An example of cepstral analysis in the spectral domain by Noll (1967:296). (a) -- logarithm power spectrum for a voiced speech segment which displays the combined effects of the peiodic source and the filter resonances. The "periodicity" of the log spectrum is seen as a regularly spaced "high-frequency ripple" in the frequency domain. (b) -- the cepstrum of the log spectrum in (a). The periodicity information in the log spectrum has been focused into a sharp peak in the high-quefrency range of the cepstrum. The broader low-quefrency peaks are composed of the remaining resonances seen in the log spectrum.

cepstrum and the abscissa is read as changes in quefrency (Bogert, Healy and Tukey 1963). Quefrencies are expressed in units of seconds since they are equivalent to cycles per Hertz (Noll 1967). The corresponding cepstrum for the log spectrum in Figure 2.14a is shown in Figure 2.14b. The cepstral analysis focuses the periodicity information in the log spectrum of the input speech into a sharp high-quefrency peak. The broader low-quefrency peaks are composed of the remaining resonances seen in the log spectrum.

Noll (1967) presented an early scheme for pitch period detection using the cepstral analysis method. As can be seen in Figure 2.14b, a sharp peak is found in the cepstrum at the pitch period of the analyzed speech segment. No clear peak is evidenced for a cepstrum derived from a segment of unvoiced speech. Thus, the cepstrum may be used to make voiced/voiceless decisions as well as estimating the pitch period of voiced speech sounds. A windowed segment of speech is transformed into the cepstrum by an inverse Fourier analysis of its log spectrum. A threshold method is used to determine the presence and location of a significant cepstral peak in a pre-designated high-quefrency range. If a peak is detected then the speech segment is classified as voiced and its quefrency measure translated into F0. The cepstrum is favorably weighted towards the high-quefrency end to improve peak detection since cepstral peaks tend to decrease in amplitude as quefrency increases. Additional logic is used by Noll to correct local extraction errors, in particular, pitch period doubling associated with the appearance of a second "rahmonic" in the cepstrum. As Hess (1983) pointed out, the usefulness of the cepstral analysis for pitch detection is actually due to the logarithmic transformation of the power spectrum

rather than its deconvolution (which is more useful for cepstrally smoothing the log spectrum). The logarithm serves as a form of spectral flattening which improves the focusing of periodicity information by the inverse Fourier analysis. Figure 2.15 (from Noll 1967:298) is a series of log spectra and cepstra derived from segments of continuous speech produced by a male speaker. Each frame from top to bottom represents a 40 ms segment of speech, analyzed at 10 ms intervals. It appears that the first seven sections from the top are voiceless since no clearly defined high-quefrency peaks are seen in these cepstra. In the remaining cepstra, there are cepstral peaks which increase in quefrency over time. That is, the speech segment evidenced a fundamental frequency which decreased with time.

Rabiner and Schafer (1978:377-378) mentioned some difficulties which might arise when using the cepstrum for pitch detection. Firstly, the lack of a large peak in the cepstrum does not necessarily mean that the speech segment is unvoiced. The cepstral method requires at least two pitch periods in the analysis segment to prevent undefined pitch peaks. Therefore, large speech intervals are required for cepstral analysis in the case of low pitched speakers. If a segment under analysis is too large then the notion of quasi-stationarity of the parameters will not hold. Rabiner and Schafer (1978:378) suggested a variable sized speech analysis segment which could be adapted to the appropriate length depending on the expected pitch period. Secondly, cepstral analysis has difficulty with extremely bandlimited speech (e.g., voiced stops) in which a minimal periodic structure is present in the log spectrum (a form of systematic failure noted by Martin (1981)). To improve

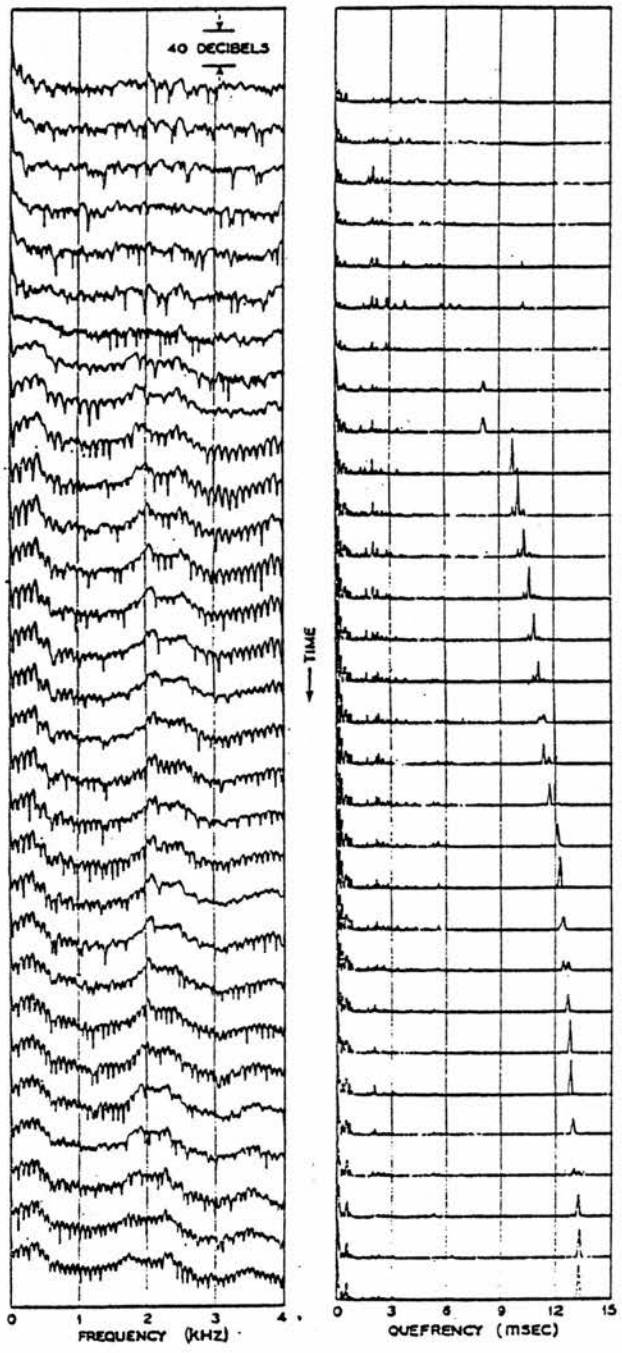


Figure 2.15 A series of log spectra (left) and their equivalent cepstra (right) derived from a sample of connected speech produced by a male speaker. Each frame from top to bottom represents a 40 ms sample of speech, analyzed at 10 ms intervals. (Figure from Noll 1967:298).

pitch period detection by cepstrum, Rabiner and Schafer suggested the addition of zero-crossing counts and energy information as additional measures of confidence.

Harmonic Analysis

This section describes a variety of techniques for direct examination of the short-term frequency spectrum of voiced speech for the fundamental frequency. The most direct measurement of F0 is the detection of the spectral peak associated with the F0 of the short-term spectrum. However, the success of this method cannot be guaranteed since the F0 spectral peak may not be present in the spectrum (as in bandlimited speech, for example) or weak in amplitude. Direct spectral peak detection is also limited by the accuracy of frequency measures in the low-frequency range of the spectrum. Other methods for detecting pitch information in the spectrum simulate visual detection of F0 in speech spectrograms. In one method, F0 is treated as a fraction of the higher harmonics, that is, F0 as the lowest common divisor of the harmonics present in the spectrum. This treatment leads to pitch extraction techniques known as spectral compression and harmonic matching. The other method observes that F0 is the distance between adjacent spectral peaks in the spectrum. The major advantage of the harmonic analysis extractors is that F0 need not be present in the speech spectrum. Further, harmonics are fairly noise-resistant and amenable to increases in frequency resolution by interpolation techniques.

Spectral Compression — Schroeder (1968) proposed a number of pitch extraction techniques based on the method of spectral compression. Spectral compression techniques rely on the condition that if a given spectrum contains harmonics of the F0 then compression of the spectral frequency range by whole number divisors will highlight the F0. The problem of spectral compression was stated by Schroeder (1968:830) as follows:

> '...admit all integer submultiples of all measured frequencies as possible values of the fundamental frequency and then select the correct value of the fundamental frequency by searching for coincidences (or near coincidences) amongst these submultiples.'

The advantage of spectral compression procedure is that the actual harmonic numbers of the spectral peaks do not need to be known.

Schroeder's first proposal was the use of a frequency histogram to spectrally compress a spectrum (see Figure 2.16, from Schroeder 1968:830). Bandpass filters spanning selected sections of the entire frequency range of interest for pitch detection are applied to the speech segment under analysis. The filtering resolves the input signal into its harmonic components. A histogram is created in which frequencies of the output signals of the bandpass filters are the entries (i.e. the bins). The harmonics are then spectrally compressed by successive divisions by whole number multiples and the resultant frequencies are entered into the histogram (e.g. compression by factors of 2, 3, 4 etc.). A maximum appears in the frequency histogram at a frequency bin representing the F0. In the Figure, note that four entries are present in the f1 bin denoting the number of trials (i.e. compressions) as well as the maximum of the function. The inverse technique known as the period histogram was suggested by Schroeder as an easier histogram method to

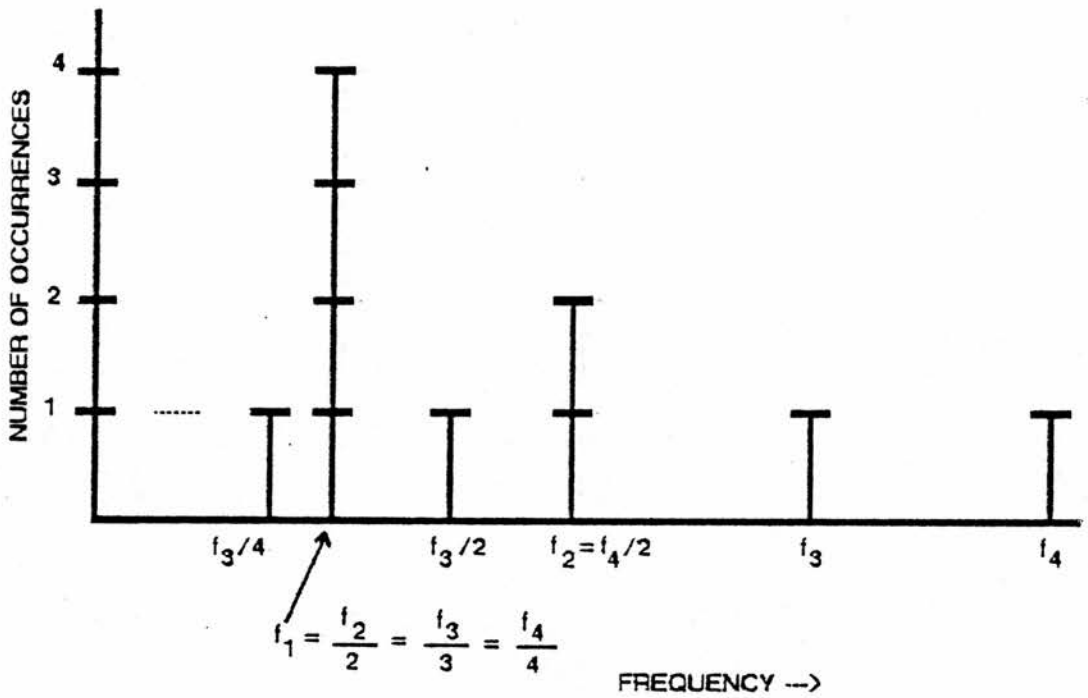


Figure 2.16 An example of spectral compression harmonic analysis using the frequency histogram technique of Schroeder (1968:830). The input signal was bandpass filtered to produce the fundamental, second, third and fourth harmonic frequencies. For each harmonic, an entry is placed in the appropriate bin for the frequency value fn. The frequency values are compressed by successive division and the resultant subharmonics are registered in the histogram (4 compressions in this example). Note that a maximum occurs in the bin f1 which represents the fundamental frequency component.

implement in hardware since multiplications could be used to fill the bins. In addition, weighting of the histogram entries by spectral peak amplitudes could provide more accurate and reliable F0 measures. In this case, harmonics which are high in amplitude and less corrupted by noise will be favorably weighted during the spectral compression. Schroeder found that the histogram techniques of spectral compression produced equivalent results to cepstral pitch analysis without the need for an additional Fourier analysis.

The histogram technique for spectral compression of harmonics was generalized by Schroeder (1968) and Noll (1970) to methods which compress all data present in a short-term spectrum. In the generalized case, all spectral frequency values are divided by a whole number and then added to the original frequency spectrum -- larger and larger divisors are used until the summated spectrum reveals a maximum at a location equivalent to the F0. This type of spectral compression causes spectral harmonic information to add coherently while non-periodic spectral information adds non-coherently. Figure 2.17 (from Noll 1970:784) is an example of compressing the log spectrum of a voiced sound. Part 1 displays the original frequency spectrum of the input signal while parts 2 and 3 show the spectra which result when the original spectral values are compressed in frequency by factors of 2 and 3. Part 4 depicts the resultant summation of the three spectra -- note how a spectral peak located at the F0 is emphasized by this technique. Taking the antilog of the compressed and summated log spectrum produces the harmonic product spectrum. This compression procedure has the advantage of spectral weighting since the actual log amplitude values are used in the summation. Noll's version of the harmonic

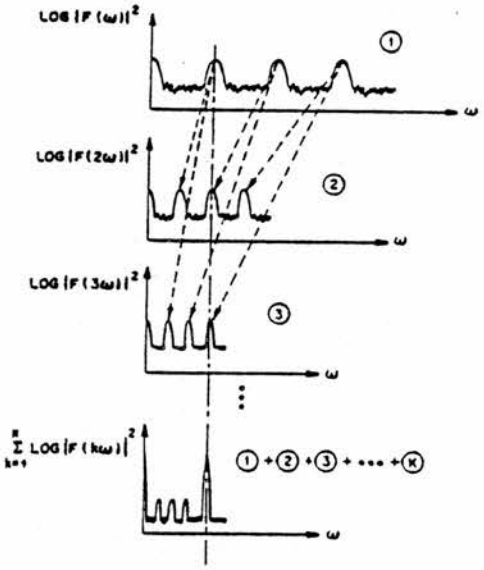


Figure 2.17 Generalization of the spectral compression technique for pitch detection in the spectral domain. In this harmonic analysis detector, all frequency components of a log spectrum (part 1) are compressed by whole multiples (parts 2 and 3) -- the resultant compressed spectra are summated together (part 4) which emphasizes the energy at the fundamental frequency region. (Figure from Noll 1970: 784).

product spectrum operated on speech data which was spectrally flattened by a Hamming window prior to the compression procedure. Figure 2.18 (from Noll 1970:785) displays a series of spectra created by compression techniques for a segment of continuous speech produced by a female speaker. The left column is a series of log harmonic product spectra displayed in time from top to bottom. The right column displays the harmonic product spectra (i.e. the antilog equivalents of the spectra in the left column). Each harmonic spectrum was created by five compressions of the original input spectrum. Note the clear F0 peak in each harmonic spectrum. Noll (1970) investigated the harmonic product spectra since it was found that the cepstral pitch extraction technique did not work very well for noisy signals despite the appearance of a harmonic structure in the spectra of these waveforms. Cepstral and harmonic compression techniques for pitch detection were compared by Noll for noisy and regular speech signals. The results of the comparison demonstrated that the harmonic product compression technique produced better pitch measures for synthetic speech signals in the presence of noise (0 dB S/N) than the cepstral method. However, Noll (1970:793) concluded that cepstral extraction performed "better" for speech data free of noise.

Harmonic Matching -- The spectral compression for pitch detection is useful since the compression procedure focuses all pitch information in a spectrum into a maximum where F0 is represented as the least common divisor of all its higher harmonics. Pitch detection by harmonic matching is a more synthetic approach in that idealized spectra of known harmonic structure are matched to a given input speech spectrum until a best fit is found between the

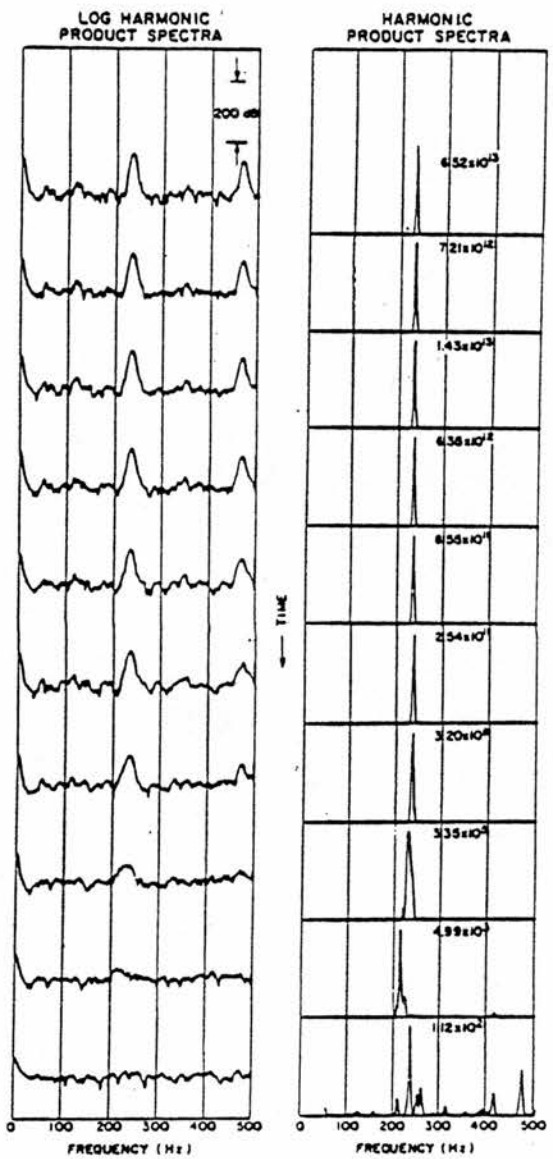


Figure 2.18 Examples of spectral compression for pitch detection in the spectral domain from Noll (1970:785). A series of compressed spectra are shown from top to bottom of this figure; the amplitudes of the spectra have been normalized before plotting. Left column — log harmonic product spectra (as produced by the technique shown in Fig. 2.17); right column — harmonic product spectra (the antilog equivalents of the spectra in the left column). Note the clear peaks in the harmonic product spectra at the fundamental frequency.

two. The essential features of a harmonic matching F0 detector are 1) an input speech spectrum with F0 to be extracted (this F0 may or may not be present in the spectrum), 2) a generated set of harmonic spectral templates for comparison to the input function and 3) a logical decision algorithm for choosing the best match between a given harmonic template and the input spectrum. In most detectors of the harmonic matching type, the input spectrum is created by Fourier transformation of the speech signal while the harmonic templates are in the form of a spectral comb function. The spectral comb usually takes the shape of an impulse sequence (the "teeth") with an intra-tooth distance equal to the trial F0. The basic differences between harmonic matching algorithms are the ways in which input and generated spectral functions are manipulated to improve the matching process. Also, the decision logic for determining the best harmonic fit differs from extractor to extractor.

The detector created by Martin (1981; 1982) exemplifies all the basic concepts of pitch detection by harmonic matching. A series of trial spectral comb filters (with F0s in a given range of interest) are crosscorrelated with an amplitude spectrum derived by a short-term Fourier transformation of a window of input speech. The maximum correlation between input and trial spectra is the best fit and the F0 is taken from the comb function (the best fit procedure is equivalent to measuring the amount of energy from the input spectrum which passes through a given comb filter). The input speech spectrum is manipulated in a manner which limits faulty F0 measures due to the presence of intra-harmonic noise. This noise may be due to background noise, non-stationary effects such as pitch

period perturbation, truncation of the speech signal by windowing, etc. and appears as non-zero spectral values between harmonic peaks. To suppress the intra-harmonic noise, an amplitude threshold level is chosen which only permits spectral peaks of substantial energy to remain in the spectrum; the threshold is an adaptive one based on the maximum amplitude of the spectrum. Each detected peak in the spectrum is defined as its maximum spectral value plus the two neighboring values — this definition reflects the fact that spectral peak frequencies may not be exact harmonic multiples of the F0 due to computational and phase distortions caused during Fourier analysis. Parabolic interpolation is applied to each detected peak to determine the frequency location and, as a consequence of this procedure, a considerable improvement in frequency resolution is also provided. The input spectrum is now comprised of highly resolved significant peak frequencies with all other spectral samples set to zero. The spectral comb function is not flat in amplitude -- the amplitudes of the teeth are weighted to prevent sub-harmonic or second harmonic tracking due to overemphasizing the high-frequency components in the spectrum (Hess 1983). In Martin's system, the teeth of the comb function decrease in amplitude as frequency increases, and the exact spectral slope is determined by a given weighting function. A trial comb filter is applied to the modified input spectrum to produce a crosscorrelation estimate. A series of trial comb functions are applied to the speech spectrum in this manner for a F0 range of interest. The resultant crosscorrelation function demonstrates a maximum for the best fit between comb function and input spectrum and the F0 is taken from that comb filter. Martin (1982:182) reported that his harmonic matching detector worked better than the cepstral technique due to

certain systematic failures of cepstral analysis. These failures of cepstral analysis occur when only the F0 or two consecutive harmonic are present in the input spectrum.

Most of the harmonic matching detectors are variations on the basic features found in Martin's system. Paliwal and Rao (1983) match an input speech spectrum with comb filters weighted by a smoothed spectrum. This comb filter weighting is produced by LPC filter analysis of the input spectrum and serves as a spectral flattener. The best fit between input spectrum and weighted comb functions occurs when a minimum difference is found (this method is similar to the minimum difference approach of the AMDF). Duifhuis, Willems and Sluyter (1982) presented a harmonic matching detector which is based on Goldstein's (1973) psychoacoustic model of pitch perception of complex signals. The input speech spectrum is searched for spectral peaks that fulfill certain psychoacoustic assumptions such as minimum audibility, masking of tones by noise, and auditory insensitivity to phase and amplitude for pitch perception. The trial functions which are compared to the input spectrum are in the form of a harmonic sieve which passes significant spectral peaks at given harmonic intervals (similar to Martin's 1981,1982 comb filter). The harmonic sieve is designed with tolerance intervals at the filtering points which are frequency dependent. A minimum squared error criterion is used as a minimum distance measure to determine the best fit of a given sieve to the input spectrum. Terhardt, Stoll and Seewann (1982a; 1982b) also presented a pitch extraction algorithm based on rigorous psychoacoustic principles as developed by Terhardt (1974). The final decision logic of this system uses subharmonic matching

principles in a manner similar to a histogram approach (Hess 1983). Sreenivas and Rao (1979) developed a harmonic matching system with the aim of detecting F0 in noisy conditions. This matching algorithm depends on searching the input spectrum for a small number of spectral peaks which are clearly defined. Thus, the matching routine is provided with strong indicators of harmonic structure from which F0 may be determined in the presence of noise.

F0 as the Distance Between Adjacent Spectral Peaks -- In this harmonic analysis technique, the fundamental frequency is measured as the distance between adjacent peaks in the frequency spectrum where these peaks represent the higher harmonics of the quasi-periodic speech signal. One recent example of a PDA operating in this manner was presented by Seneff (1978). Though not discussed in detail here, one of the main aspects of this particular PDA is the use of a variety of techniques to greatly reduce the computational effort usually associated with the spectral analysis of speech for the F0. The first step in the spectral peak analysis is the elimination of redundant and irrelevant information from the input speech signal. The signal to be transformed by an FFT need only contain a 700 Hz bandwidth which permits a maximum adjacent peak distance of 350 Hz (i.e. the highest permitted F0). In addition, this PDA was developed for speech transmitted via a telephone line and therefore frequencies below 300 Hz are considered irrelevant. The following steps are completed to produce the new signal:

1) The speech signal is digitized at a sampling rate of 7.5 KHz to obtain a digital signal with frequencies up to approximately 3.75 KHz.

2) The frequency content of the waveform is limited by downsampling the digital signal by a factor of 3 (the

signal is first low-pass filtered with a cutoff set at 1.26 KHz). The downsampling produces a signal s1(n) containing frequencies from -1.26 KHz to 1.26 KHz. The downsampling procedure does not decrease the accuracy of measurement for frequency domain PDAs as it does for time domain PDAs (Hess 1983).

3) A spectral rotation method is used to rotate the spectrum of s1(n) by 90 degrees in its z-plane representation. The spectral rotation produces a signal s2(n) in which the origin of the original signal has been rotated to 630 Hz. See Seneff (1978) for a detailed explanation of the spectral rotation method.

4) The downsampling procedure of step 2 above is applied to the spectrally-rotated signal to produce a complex waveform s3(n) containing frequencies up to 420 Hz. According to the Seneff (1978:359):

'...because of the rotated spectrum, 630 Hz in the original waveform corresponds to zero Hz in s2(n), and thus our doubly downsampled complex waveform contains the information from (630-420)Hz to (630+420)Hz in the original speech, which is the desired spectral region.'

The new signal s3(n) is then transformed to the frequency domain by an FFT. The basic extractor applies a peak-picking procedure to the spectral peaks in order to determine the correct F0. Prior to the peak-picking routine, irrelevant and/or erroneous peaks are excluded from the analysis based on the lack of an appropriate amplitude or spectral distance in relation to other adjacent peaks. A rank-ordering of the remaining spectral peaks is used to fill a table with F0 estimates. An iterative approach is used to rank-order the peaks according to their amplitude (i.e. by successively decreasing peak amplitudes) with each newly detected peak used to determine the spectral distances to its neighbors. These iterations continue until a certain number of F0 estimates are found within a given small tolerance interval or no more peaks are found within the spectrum. This basic extractor has features similar to the above frequency histogram techniques as well as to the channel selection routine of Gold and Rabiner (1969).

SECTION 2.2.3 -- SUMMARY -- SPECTRAL DOMAIN PDAs

A brief review of spectral domain pitch detection algorithms was presented in the preceding sections of this chapter. The chief characteristic of spectral domain PDAs is the short-term transformation of the input time domain waveform to some other spectral domain -- the transformation is completed by the pre-processor component of these PDAs. Useful short-term transformations for pitch detection in the spectral domain are those which concentrate all the periodicity information contained within a frame of voiced speech into a single peak value. The basic extractor of the spectral domain PDAs determines the location and strength of the peak value. The use of short-term transformation techniques means that the resultant period or F0 value detected by a PDA represents an average estimate of pitch for a given frame of voiced speech. Two general modes of pitch detection in the spectral domain were discussed here. One mode uses correlation functions to transform the input speech waveform to a lag domain representation of the signal. This type of transformation produces a function which displays time as a delay between two correlated input signals. The autocorrelation and average magnitude difference functions have been implemented as part of a number of spectral domain PDAs. The other mode of spectral domain pitch extraction relies on a transformation of the speech signal to a frequency domain representation. The short-term transformation of the signal to the frequency domain is usually completed by a series of bandpass filters spread across the frequency range of interest or by Fourier analysis. A number of frequency domain PDAs directly examine the harmonic structure of a speech signal by spectral compression or

harmonic matching techniques to produce measures of period or F0. Cepstral analysis consists of multiple spectral transformations of the input signal to produce a time-based cepstrum from which period data can be extracted.

The main advantage of pitch extraction in the spectral domain is the resistance of these PDAs to corruption of the input signal by additive noise. Spectral domain PDAs are particularly useful for extracting pitch data from bandlimited signals in which the fundamental harmonic component may be absent. The main disadvantages of these PDAs is the computational expense associated with the various spectral transformation techniques and the smearing of cycle-to-cycle information by the short-term transformation of the speech signal.

CHAPTER 3

THE MODIFIED PARALLEL PROCESSOR FOR EXTRACTING FUNDAMENTAL FREQUENCY AND AMPLITUDE CONTOURS FROM THE TIME DOMAIN REPRESENTATION OF CONNECTED SPEECH

CHAPTER 3

THE MODIFIED PARALLEL PROCESSOR FOR EXTRACTING FUNDAMENTAL FREQUENCY AND AMPLITUDE CONTOURS FROM THE TIME DOMAIN REPRESENTATION OF CONNECTED SPEECH

3.0 INTRODUCTION

The automatic system for the acoustic analysis of waveform perturbations in connected speech consists of three major components including 1) a pitch detection algorithm for extracting fundamental frequency and amplitude contours from a time domain representation of the speech waveform, 2) a non-linear smoothing algorithm for determining the smoothed trend lines of the fundamental frequency and amplitude data and 3) statistical evaluation of the deviations between the original contours and their associated smooth trend lines which produces the intonation and perturbation parameters. This chapter focuses on the first element of the system in which an input speech waveform is analyzed for F0 and A0 data by a parallel processing pitch detection algorithm operating in the time domain. The point of departure for this algorithm is one devised originally by Gold and Rabiner (1969).

As Hiller et al. (1983) noted, one of the problems in choosing a suitable method for the automatic detection of the duration of pitch periods in the acoustic waveform is that there is a tension between two quite different needs: the need to establish the smoothed trend which represents the intonational contour, versus the need to register as accurately as possible the momentary deviations (or 'excursions') of individual periods from this smoothed trend,

representing phonatory quality. Most pitch detection algorithms involve a good deal of smoothing in their inherent design, and as such are well-suited to gathering intonational data. There are very few algorithms available that are capable of tracking the exact durations, cycle by cycle, of the perturbed train of periods that is characteristic of not only dysphonic, pathological voices, but also of many types of healthy voices. The parallel processing detector was chosen in the light of criteria emerging from comparative studies of a number of pitch period detection algorithms (Rabiner et al. 1976; McGonegal, Rabiner and Rosenberg 1977; see appendix for the paper by Laver et al. 1982). The Gold and Rabiner method was felt suitable for the present study's needs in that it can work on connected speech from both male and female speakers and retains accuracy of period duration estimation in conditions of fairly acute waveform perturbation in both fundamental frequency and amplitude. As discussed in Chapter 2 above (Section 2.1.2), this type of PDA is considered to be a very powerful time domain detector without being the most complicated — the parallel processor's power is mainly due to its multichannel approach to the structural analysis of the input speech waveform.

The discussion in this chapter begins with detailed descriptions of the algorithms which comprise the modified parallel processor for fundamental frequency and amplitude extraction in the time domain. The pre-processing of the input speech signal prior to pitch extraction is also discussed at this point. The appropriate analysis conditions for obtaining frequency and amplitude data via the parallel processing PDA are then presented. Based on these analysis conditions, a small study is reviewed in which the

performance of the parallel processor is compared to visual examination of speech waveforms for pitch extraction. A section then follows in which the issues of sampling resolution and interpolation are discussed in relation to pitch detection in the time domain. Finally, the range of applications of the parallel processor is considered.

SECTION 3.1 — PARALLEL PROCESSING FOR DETECTING PITCH PERIODS IN THE TIME DOMAIN

The basic scheme of the parallel processor, as a very fast program implementable on a general purpose computer, has been described by Rabiner and Schafer (1978:136) as follows:

> '1. The speech signal is processed so as to create a number of impulse trains which retain the periodicity of the original signal and discard features which are irrelevant to the pitch detection process.
>
> 2. This processing permits very simple pitch detectors to be used to estimate the periodicity of each impulse train.
>
> 3. The estimates of several of these simple pitch detectors are logically combined to infer the period of the speech waveform.'

The idea of parallelism in period detection is that the outputs of a number of simple parallel measures of periodicity for a given speech segment are the inputs to a sophisticated matching process based on majority logic which determines the segment's most likely pitch period. Gold and Rabiner (1969) suggested that parallelism, as implemented in an automatic pitch period estimator, may simulate the visual observations of a human examining a speech waveform for periodicity (although this suggestion must be regarded as an over-simplification and/or as an untested hypothesis about the

functioning of the parallel processor). The parallel processor analyzes the overall temporal structure of a speech waveform to determine pitch periods.

A block diagram of the parallel processor is shown in Figure 3.1 (adapted from Gold and Rabiner 1969). The input speech waveform is pre-processed in two stages including 1) phase compensation of low-frequency phase distortions caused by the tape recording process (a modification of the system not presented as part of the original Gold and Rabiner algorithm) and 2) low-pass filtering to eliminate high-frequency harmonic components, which simplifies the signal structure. The basic extractor of the parallel processor consists of several parts. Firstly, the low-pass filtered speech is processed to produce several functions representing different aspects of periodicity in the waveform. A simple primary extractor is then applied to each function to determine the periodicity displayed by that function. A matching process combines the various measures of periodicity to determine the duration of the pitch period for the input signal. In addition, processes are required for determining the presence of speech (i.e. discrimination between speech and silence) as well as the likelihood of the resultant pitch period representing a voiced or unvoiced segment. An elaborate post-processing routine is not required for the output of the parallel processor except for the conversion of period measures to units of Hz. The general structure of the program follows the more elaborate version of Gold and Rabiner's (1969) parallel processor in order to accommodate the widest variety of voice types. In the present implementation, the program completes the parallel processing of a given window of speech data and then the window is

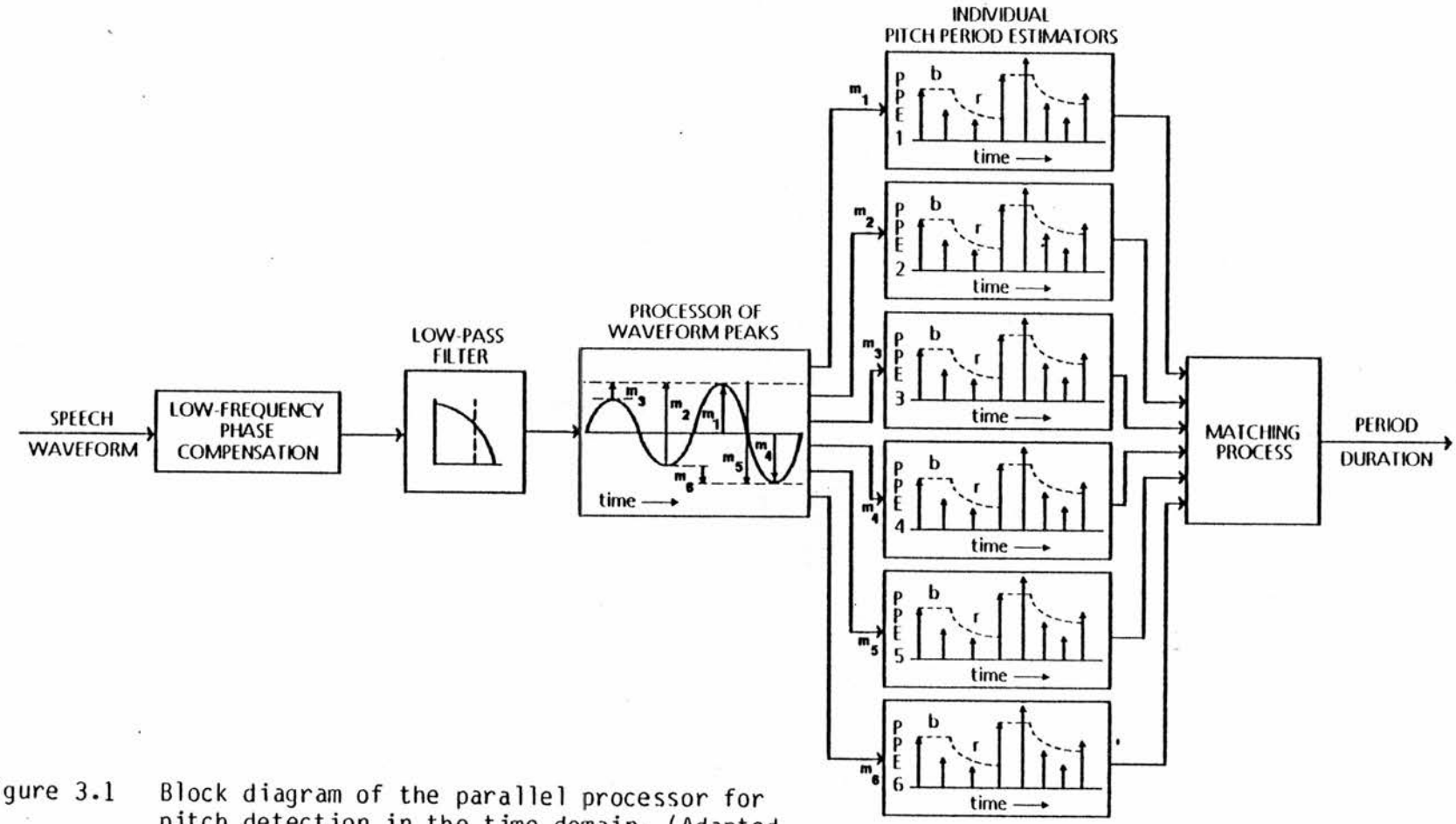


Figure 3.1 Block diagram of the parallel processor for pitch detection in the time domain. (Adapted from Gold and Rabiner 1969).

shifted forward in time to try to capture the next pitch period.

The parallel processor processes a number of useful characteristics for detecting pitch periods in voice samples elicited from healthy and pathological speakers — this data being the input to an algorithm which evaluates waveform perturbations. The parallel processor operates in the time domain by analyzing the general temporal features of an input signal for indicators of waveform periodicity. Like other time domain PDAs, the parallel processor completes a local analysis in which individual pitch periods are detected directly from the time domain signal (most spectral domain PDAs perform a global analysis which produces an average measure of period or F0 for a small number of pitch periods). The advantage of this characteristic is the ability of the time domain PDA to track rapid changes of F0 even in slightly irregular signals (Hess 1983). A local analysis of pitch is required for a perturbation measurement system since waveform perturbations are by definition measures of cycle-to-cycle fluctuations of period and amplitude. Of course, the ability to track pitch periods in irregular signals is most useful when analyzing the speech of pathological speakers. Hess (1983) described the parallel processor as "fault-tolerant" since much of the program logic is concerned with the possibility of committing errors in period marker detection in the initial stages of the basic extraction process. This built-in fault-tolerance is a particularly useful characteristic when deriving data from perturbed signals. The parallel processor has been shown to track F0 data at accuracy levels similar to a number of other well-known detectors in objective (Rabiner et al. 1976) and subjective (McGonegal et al.

1977) comparative evaluations of PDAs. Fundamental frequencies ranging from approximately 50 to 600 Hz can be detected by the parallel processer (Tucker and Bates (1978) implemented a version of the Gold and Rabiner algorithm for analyzing music which could track F0 over a 6 octave range from 40 Hz to 2.5 KHz). This range is considerably wider than required for the present study in which F0 will be analyzed in the speech of male and female adults though it is clear that voice samples elicited from children could also be evaluated for waveform perturbations. Finally, Gold and Rabiner found that their algorithm can operate effectively even when the background noise level is very high (though it should be noted that this characteristic was found in an informal experiment in which no quantitative measurements were presented).

SECTION 3.1.1 -- TAPE RECORDER AND ACOUSTIC ENVIRONMENT CONTROL FACTORS

For the present study, each sample of connected speech which is input to the perturbation measurement system has been recorded on standard magnetic tape. The quality of the recording equipment and acoustic environment in which each voice sample was recorded will affect the performance of the measurement system. In particular, corruption of the speech signal by noise will adversely effect the operation of the parallel processing PDA, since it directly examines the temporal structure of the input waveform in order to extract pitch periods. Therefore, it is desired that each voice sample should be recorded under good sound quality conditions. Each sample of connected speech to be analyzed for perturbation data has been recorded on high quality tape recorders (REVOX A77 and UHER 4000)

which display very wide and flat frequency responses at a recording speed of 7.5 ips. Each voice sample is played back for digitization via similar equipment. A 70 Hz triangular wave was used as a calibration tone on the tapes to check tape recording speeds and phase distortions.

Three different recording environments were used for collecting voice samples analyzed in the present study. The departmental sound-treated recording studio in the Phonetics Laboratory at Edinburgh University was used to record the healthy control speakers. The recordings made in the studio are of a very high quality with good signal-to-noise conditions (no specifications are available for the actual signal-to-noise levels). The pathological speakers used in this study were recorded in the outpatient voice clinics of 2 hospitals. A minority of the pathological speakers were recorded in a voice clinic (Royal Infirmary Edinburgh) which had a sound-treated booth available. These recordings can also be described as high in quality with good signal-to-noise conditions. The remaining pathological speakers were recorded in a clinic (Radcliffe Infirmary Oxford) in which the conditions varied from recording to recording. Some of the pathological speakers were recorded in a sound-treated booth which resulted in high quality recordings with good signal-to-noise conditions. However, the majority of the pathological speakers were recorded in a therapy room which had not been treated for sound. The recordings made in these conditions contain both continuous (e.g. air conditioning fans) and transient noises (e.g. closing doors, typewriters, etc.). On some voice samples, electrical hum from the mains electricity was evidenced as a 50 Hz modulation of the speech signal. Auditory and

oscillographic examinations of these recordings were completed to determine if the the speech signal had been corrupted to the point that pitch detection would be adversely effected by the noise. Recordings which were judged as being substantially corrupted were not analyzed as part of this study. It should be noted that early in this study, quite a few recordings of pathological speakers had to be eliminated due to poor signal-to-noise conditions on the tapes. The effects of background noise on pitch period extraction in the time domain have not been formally tested in this study. Future research should investigate the effects of background noise if a perturbation measurement system such as the present one is to be applied to voice samples made in a variety of recording environments.

SECTION 3.1.2 -- ANALOG-TO-DIGITAL CONVERSION

The pitch detection process begins with the conversion of a tape recorded voice sample from an analog to a digital waveform. Each voice sample is digitized on a PDP 11/40 minicomputer at a sampling rate of 20 KHz with a 12-bit quantization of each data point. A 20 KHz sampling rate provides a resolution of .05 ms for each pitch period detected by the parallel processor. In Section 3.4 below, a rationale is set out for this choice of sampling rate, which provides a reasonable resolution of typical fundamental frequencies produced by male and female adult speakers. Prior to digitization, the input speech signal is passed through an analog filter which prevents aliasing of the waveform during sampling. The analog filter is a Butterworth type which produces a -48 dB/octave rolloff beyond a stopband frequency of 10 KHz. The digitized signal

is then filed on a VAX 11/750 minicomputer for further signal processing.

SECTION 3.1.3 -- PHASE COMPENSATION OF RECORDER DISTORTION

The first pre-processing step in the pitch period detection process is the phase compensation of low-frequency distortions of the input voice sample -- these phase distortions are the result of reactive and resistive components in the tape recording and playback system which do not maintain the relative phases of the harmonics of the recorded signal (Olsen 1982). A more accurate representation of the original speech signal is produced by compensation of the low-frequency phase distortions displayed in the recorded (and digitized) version of the signal. This accurate waveform representation is required since it was found that low-frequency phase distortions of tape recorded voice samples adversely effects the discrimination of healthy and pathological speakers by waveform perturbation parameters. A study investigating the effects of phase compensation on the analysis of waveform perturbations is presented in Section 6.2 below. The phase compensation technique used for pre-processing of an input signal is the frequency domain procedure of Berouti, Childers and Paige (1977). An FFT transforms a segment of an input waveform (interval length = 409.6 ms) to its associated magnitude and phase spectra and the phases in the spectrum are compensated by the addition of a compensation factor representing the total phase delay in a given recording/playback system. An inverse FFT transforms the compensated spectra back to a time domain representation of the speech signal.

SECTION 3.1.4 — LOW-PASS FILTERING

The essential pre-processing step for the correct functioning of the parallel processor is the low-pass filtering of the input speech signal. This filtering is the initial stage in a succession of data reduction procedures used to focus on periodicity features in the input time domain waveform. The low-pass filter simplifies the temporal structure of the signal by eliminating the spectral components above approximately the typical first formant region. The simplified waveform structure can then be searched for peak minima and maxima evidenced in the voiced segments of the pre-processed speech.

Gold and Rabiner note that the design of the low-pass filter and the exact choice of cutoff frequency are not critical to the performance of the parallel processor. If the first harmonic is present within the speech signal to be analyzed then the cutoff frequency may be low as long as it is situated above the frequency of the fundamental harmonic (e.g. a 600 Hz cutoff frequency was found to work well for the parallel processor). However, the cutoff frequency should not be too low (e.g. 250 Hz) since the resultant pitch data produced a rough quality when used in a vocoder speech synthesis experiment (Gold and Rabiner 1969).

In the present study, all the voice samples evidence the first harmonic. The major interest is in the fine details of the periodic behavior of voiced speech, that is, F0 and A0 waveform perturbations. Therefore, filter cutoffs similar to the frequencies set out in Gold and Rabiner (1969) are used for low-pass filtering, specifically 1) male speakers -- 600 Hz and 2) female speakers --

800 Hz. These cutoff frequencies provide useful filtering for the F0 ranges typical of the two sexes. The filter type is a digital linear-phase filter (McClellan 1975) consisting of 32 coefficients which produces an approximate -48 dB/octave rolloff beyond the stopband frequency. Figure 3.2 displays the frequency response curves for the low-pass filters used for the speech of males and females. It should be noted that no attempt is made to compensate the linear phase shift associated with this filtering process which is considered negligible for the present purposes.

SECTION 3.1.5 -- SILENCE DETECTION

The pitch period estimation begins by determining the presence of speech within a given window of input data. The silence detection technique is a simple one described by Gold (1964), in which the segment of data is searched for two samples which exceed a pre-determined 'silence' energy threshold. If the energy threshold is exceeded then the remainder of the estimation is completed, otherwise the pitch period result is set to zero and the next frame of data is processed. The silence detection threshold is determined interactively for each voice sample by calculating the peak intensity level of the background noise presented in each pre-processed voice sample. This simple silence detection method works well for the present study since there is a priori knowledge of the nature of the background noise found on the clinical recordings. A more sophisticated silence detection algorithm would be required if the parallel processor were to be used in a fully automatic manner. In instances where the background noise exceeds the silence detection threshold, the voiced/unvoiced discrimination

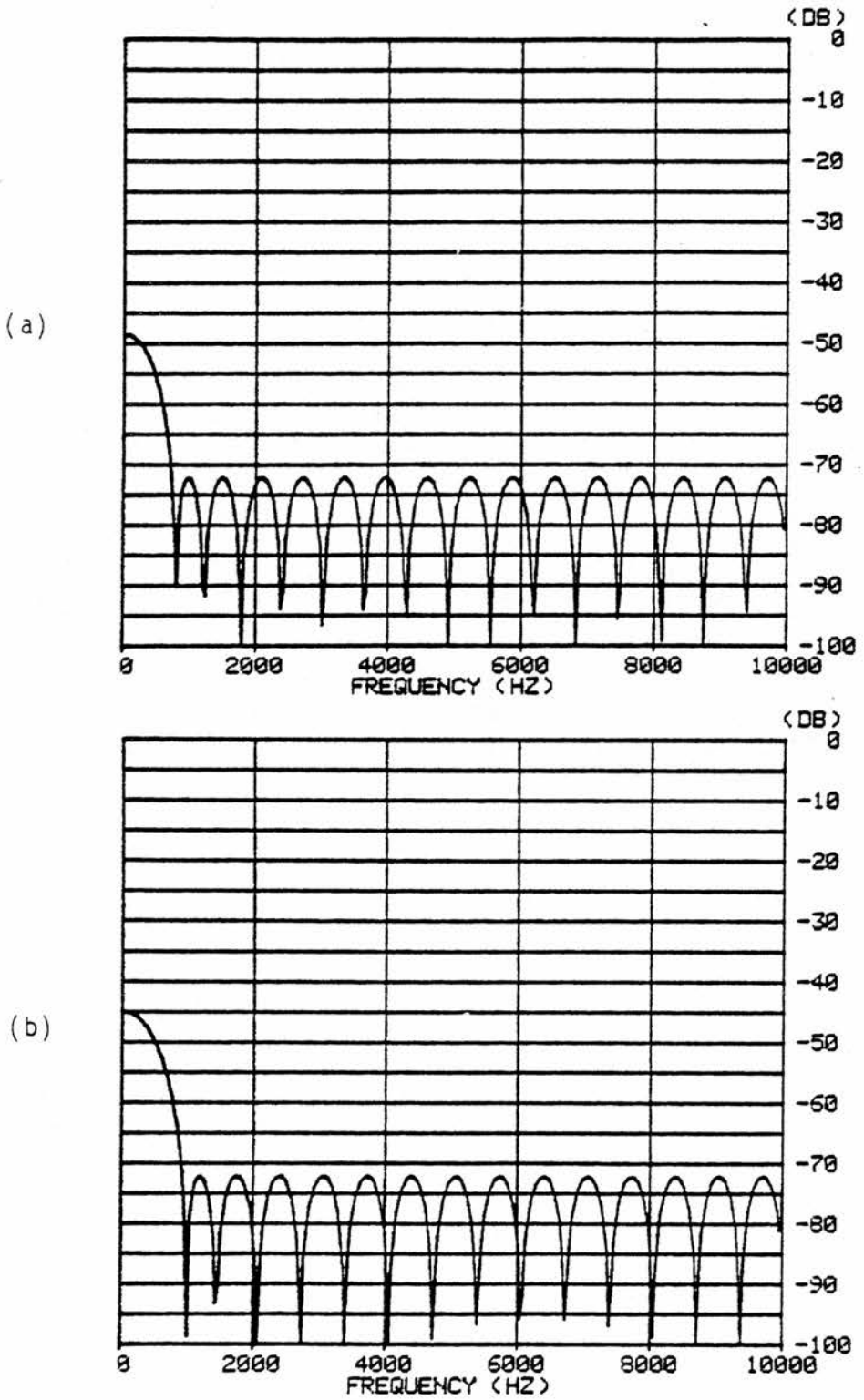


Figure 3.2 Frequency response curves of the low-pass digital filters used to pre-process speech waveforms input to the paralleling processing PDA. (a) -- Low-pass filter applied to male speech data, cutoff = 600 Hz; (b) -- Low-pass filter applied to female speech data, cutoff = 800 Hz.

algorithm (see Section 3.1.9 below) should classify the noise as unvoiced. The parallel processor does not differentiate between silence and unvoiced segments at its output and both are given zero duration pitch period values.

SECTION 3.1.6 -- DETECTION OF LOCAL PEAK MINIMA AND MAXIMA

The first stage of the basic extractor used in the parallel processor is a data reduction procedure typical of PDAs which perform temporal structural analysis of the speech waveform. This data reduction consists of selecting all waveform samples which are anchor points for determining periodicity and the elimination of the remaining data samples. In the parallel processor, peak minima ("valleys") and maxima ("peaks") are the primary anchor points for periodicity detection. The simplification of the input data to waveform peak measurements enables the use of very simple pitch period estimators in the remaining basic extraction process.

Six "functions of peakedness", labeled m1 through m6, are derived for the local minima and maxima within the pre-processed speech signal. Rabiner and Schafer (1978:137) have defined the six functions as follows:

'1. m1(n): An impulse equal to the peak amplitude occurs at the location of each peak.

2. m2(n): An impulse equal to the difference between the peak amplitude and the preceding valley amplitude occurs at each peak.

3. m3(n): An impulse equal to the difference between the peak amplitude and the preceding peak amplitude occurs at each peak. (If this difference is negative the impulse is set to zero.)

4. m4(n): An impulse equal to the negative of the amplitude at a valley occurs at each valley.

5. m5(n): An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding peak occurs at each valley.

6. m6(n): An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding local minimum occurs at each valley. (If this difference is negative the impulse is set equal to zero.)'

Figure 3.1 displays the 6 functions for a segment of filtered speech. The peak measurements are defined to occur at the same instant in time as the peaks they are associated with (i.e. m1, m2 and m3 are generated at each maximum and m4, m5 and m6 are generated at each minimum). All functions are converted into sequences of positive impulses with their respective amplitudes. In the original study by Gold and Rabiner, it is not explicitly stated if the six function definitions only apply to positive maxima and negative minima. In the present implementation of the parallel processor, this is the case and all remaining positive minima and negative maxima (whose occurrence should be fairly low in filtered speech) are also eliminated from the data structure.

The sampling resolution of each detected peak (originally sampled at 20 KHz and 12 bits per sample) is increased by parabolic interpolation of the 3 sampled data points which define the peak. See Section 3.4 below for a more complete discussion of the interpolation technique used here for increasing the sampling resolution. The increase in sampling resolution for the peaks is reached at some cost in computing effort since 1) the parabolic interpolation procedure is completed in floating point mathematics and 2) the resultant interpolated peak location and amplitude must also be stored and processed in floating point format.

The choice of these 6 peak functions is based on the possibility of processing two extreme types of signal structure including 1) a simple sinusoid (i.e. the fundamental harmonic only) and 2) a weak first harmonic combined with a strong second harmonic. Figure 3.3 (after Gold and Rabiner 1969:444) displays the 6 impulse functions for these two types of waveform. In the case of the sinusoid, measures m1, m2, m4 and m5 demonstrate strong indications of the pitch period while functions m3 and m6 fail completely. In the example where a strong second harmonic is present along with a weak fundamental harmonic, functions m3 and m6 indicate the presence of the first harmonic while the remaining functions appear to signal the second harmonic. Mechanisms exist throughout the whole of the parallel processor's basic extractor for dealing with the possibility that a majority of the functions will fail to indicate the correct pitch period as demonstrated in the above example of a weak fundamental harmonic. Since no other PDA takes into account the possibility of committing errors in the determination of period markers to such an extent, it is this which leads Hess (1983) to describe the parallel processor as the "fault-tolerant" PDA.

SECTION 3.1.7 — EXTRACTION OF PERIOD MARKERS FROM THE IMPULSE FUNCTIONS

The second stage of the basic extraction process of the parallel processor is the selection of the remaining samples which are likely to represent a period delimiter and the rejection of the other samples. This step must be completed for each of the six impulse functions based on the peak measurements. For each function, the marker detection is completed by a time-varying

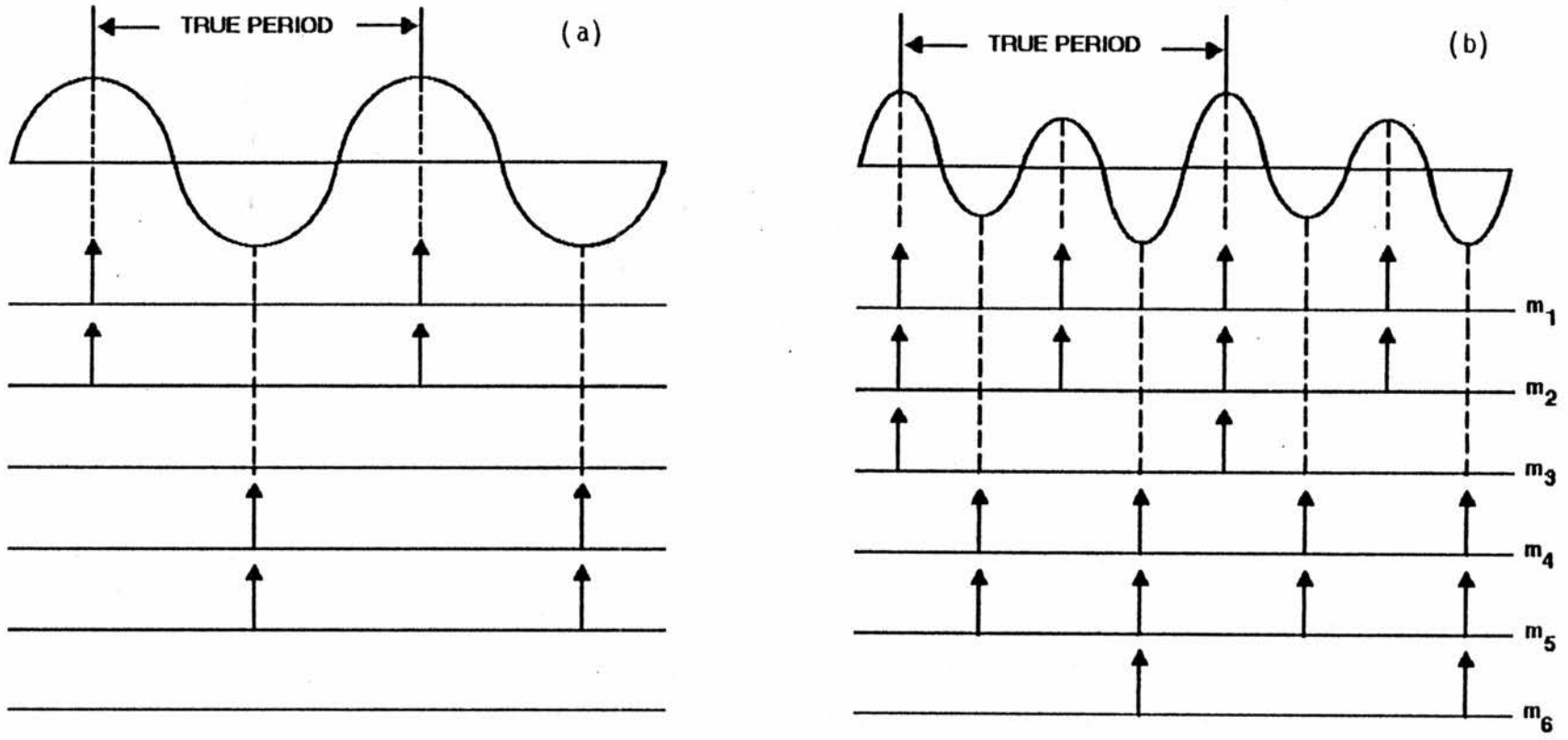


Figure 3.3 The six impulse functions produced by the parallel processor when applied to 2 extreme types of signal structure. (a) Pure sine wave -- functions m1, m2, m4 and m5 display correct period duration; (b) Weak fundamental harmonic with strong second harmonic -- functions m3 and m6 indicate the presence of the first harmonic while functions m1, m2, m4 and m5 appear to indicate the second harmonic. (After Gold and Rabiner 1969).

non-linear system (Rabiner and Schafer 1978) which consists of an exponential rundown circuit as displayed in Figure 3.1. When a pulse of sufficient amplitude is detected in the input function, its output is set to the value of the impulse. At this point, the circuit is reset in order to detect the next impulse. Upon being reset, a blanking interval occurs in which no pulse can be detected. This blanking interval prevents the tracking of strong spectral components in the pre-processed speech such as a first formant located close in frequency to the first harmonic. At the end of the blanking interval, a simple exponential decay controlled by a rundown time constant begins in order to find the next significant impulse in the function. It can be seen that the parallel processor incorporates an aspect of envelope modeling of the signal though the impulse sequence is much simplified in structure compared to the original speech waveform from which the sequence was derived. It is important to note the possibility of incorrect period detection in regions where the signal is rapidly increasing or decreasing in amplitude such that sudden changes in amplitude are registered as changes in pitch period duration. It will be shown in the following section that the parallel processor may be able to compensate for this problem but it still exists in certain extreme changes in signal amplitude. In any event, if a new pulse exceeds the level of the exponential decay, it is extracted and the rundown time circuit is reset. The result of this process is six estimates of the local pitch period based on the distances between detected impulses within each of the six impulse functions. Though not used for further pitch extraction, the peak amplitudes associated with the 6 period estimates are stored for use in the perturbation measurement algorithm.

The blanking interval (b) and the rundown time constant (r) are the time-varying aspects of the extraction process in that they are functions of a smoothed estimate of the local pitch periods detected in the speech signal. This local smoothed period duration (Pav) is based on iterative averaging as each new period is detected:

$$P_{av}(n) = \left[P_{av}(n-1) + P_{new}\right] / 2$$

where Pav is the current smoothed estimate of period duration, Pav(n-1) represents the previous estimate of pitch period and Pnew represents the most recent unsmoothed estimate of the local period. Thus, the parallel processor incorporates some local averaging in its logic in order to predict the likely duration of the next period. The values of the blanking time and rundown time constant are set as follows:

$$b = 0.4 * P_{av}$$

(i.e. no new period marker may be within less than 40% of the duration of the local smoothed average) and

$$r = P_{av} / .695$$

(this value appears to give a reasonable decay rate for impulses derived from the peaks of filtered speech). Limits are set to the actual value of Pav to prevent extremes in the values of b and r. Pav is limited to a range of from 4 to 10 ms (i.e. the smoothed local pitch period is limited to a range of 100 to 250 Hz). If the computed value of Pav exceeds one of these limits then b and r are

derived from the value of the exceeded limit.

SECTION 3.1.8 — MATCHING PROCESS FOR DETERMINING THE MOST LIKELY PERIOD ESTIMATE

The final step of the basic extraction process is the selection of the parallel channel which contains the most likely period estimate amongst the six duration estimates produced by the primary extractors. This is the major problem for any multichannel processor in that the correct channel must be selected when other competing possibly incorrect channels are also active. A matching process is used for channel selection in the parallel processor as first set out by Gold (1964).

The matching logic is designed with respect to the error characteristics found for the six basic measures of periodicity (Hess 1983). Firstly, the extraction error known as a "hole" (i.e. where an individual marker has been missed) should not occur if the blanking interval of the primary extractor does not exceed the period duration -- functions m1, m2, m4 and m5 will always produce impulses for their given polarities if the analysis interval is long enough in duration. Second, there is a tendency towards higher harmonic tracking in the parallel processor since the blanking interval and exponential decay rate are controlled by a smoothed period estimate which may not be correct in relation to the incoming pitch period. With this second factor in mind, the overall matching process uses the bias towards higher harmonic tracking in order to apply a minimum-frequency selection principle in which the channel containing the longest duration (i.e. the lowest frequency) is selected as the most likely period.

The matching process begins by forming a matrix of pitch period estimates from which the most likely period duration may be derived. Figure 3.4 (after Gold and Rabiner 1969:445) displays the matrix (a) as well as an example (b) of where some of the matrix entries have been derived from a given periodicity function. Each row of the matrix consists of the following period estimates:

> Row 1 -- The six period estimates derived by the most recent exponential detection of the six impulse functions.
>
> Rows 2 and 3 -- The previous 2 sets of period estimates derived from the six pulse trains.
>
> Rows 4 and 5 -- Row 4 is the summation of the durations from rows 1 and 2 while row 5 is the summation of the periods contained in rows 2 and 3; the purpose of rows 4 and 5 is to ensure correct period estimates when some of the first row candidates are actually measures of a second harmonic within the pulse function.
>
> Row 6 -- Row 6 is the summation of the durations contained within the first 3 rows; the purpose of this row is to prevent the tracking of the third harmonic component displayed by one or more of the primary candidates.

Figure 3.4b displays the various duration measures stored within the matrix for two of the impulse functions (PPE1 and PPE2) -- the indices of the various pitch period measures can be used to locate their positions within the matrix.

Having formed the 6x6 matrix of period durations, the matching process then searches for the most likely period duration contained within the entries. Only the 6 duration estimates of the first row are possible candidates for final selection. Each candidate is compared with the other 35 durations to determine similarities amongst the data. These durational similarities are called "coincidences" -- the first row candidate which demonstrates the greatest number of coincidences with the remaining estimates is chosen as the most likely pitch period. The measurement of

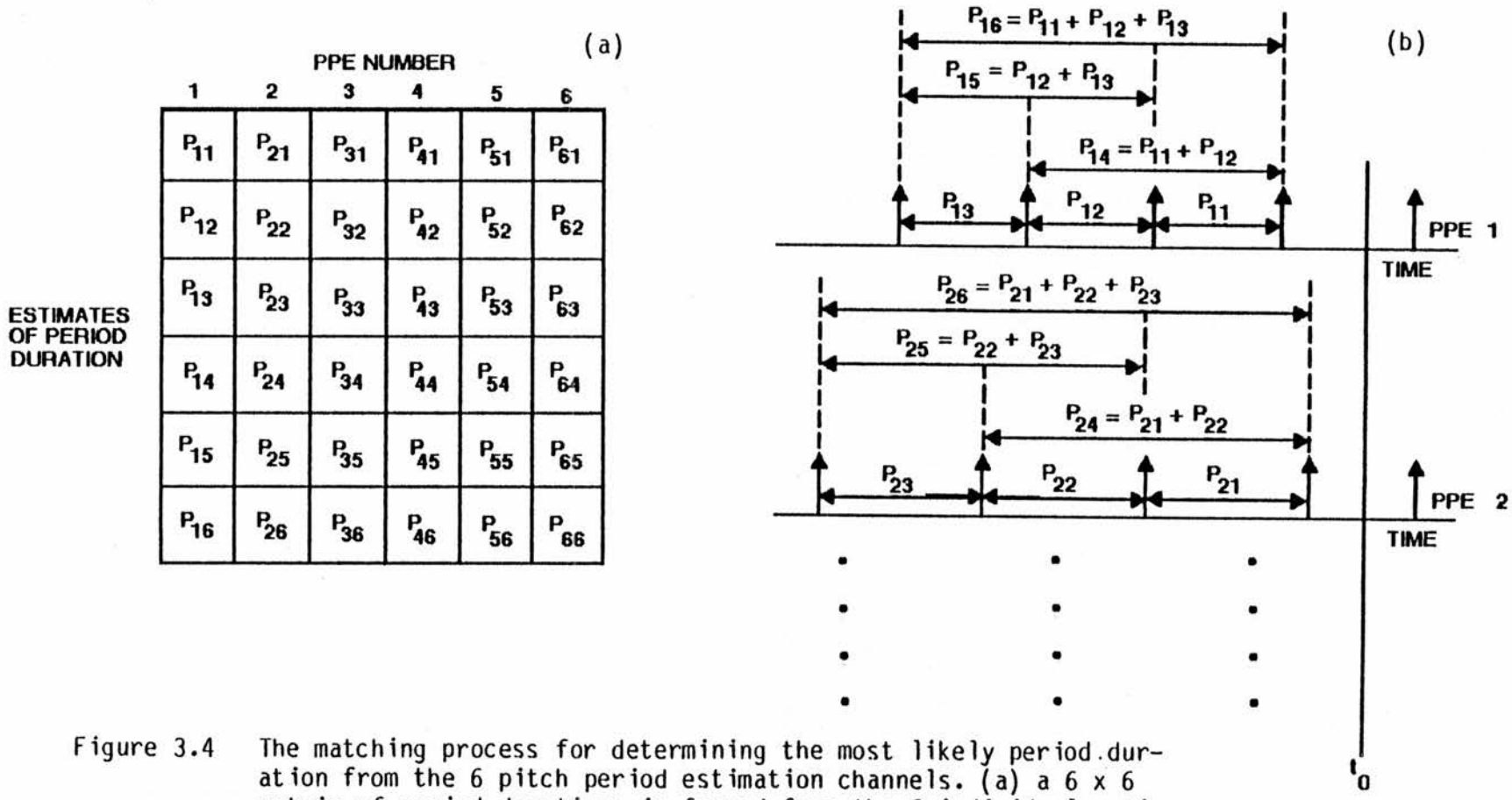


Figure 3.4 The matching process for determining the most likely period duration from the 6 pitch period estimation channels. (a) a 6 x 6 matrix of period durations is formed from the 6 individual estimates of period length (represented by the first row); (b) two examples (PPE1 and PPE2) of the computation of the period estimates used to complete the matrix. Only the estimates of the first row are possible candidates for final selection. (After Gold and Rabiner 1969).

coincidence is based on the absolute difference value found between a primary candidate and each of the other durations contained within the matrix. Due to the natural cycle-to-cycle variation evidenced in time domain speech signals, the registration of a coincidence need not be based on an exact match between the candidate and a given other estimate (i.e. an absolute difference of zero between the 2 durations) but rather on thresholds which allow some tolerances in the matching process. The thresholding scheme of the matching process is displayed in Figure 3.5 (after Gold and Rabiner 1969:445). A set of pitch period ranges (in units of ms) are listed down the left portion of the Figure. Along each row of the Figure is a set of coincidence window widths (i.e. the thresholds) for each one of the pitch period ranges. For example, a candidate period of 10 ms falls within the range specification of row 3 (i.e. 6.3 - 12.7 ms) -- an absolute difference of less than or equal to 400 μs must occur between this candidate and some other given period estimate in order for a coincidence to be registered (see column 1 of row 3). Two characteristics can be noted for the table of coincidence window widths displayed in this Figure. Firstly, looking down each column it can be seen that as a candidate period falls within larger and larger ranges (i.e. each range is a doubling of the previous one), the width of the coincidence window increases which permits greater tolerance for longer pitch periods. Secondly, there are four coincidence window widths per row for each pitch range -- each primary candidate is compared to the other 35 estimates over 4 repetitions, each time with an increasing window width, that is, each set of comparisons has decreased accuracy requirements in order for a coincidence to be registered. Biasing factors are also displayed across the top of the table in Figure

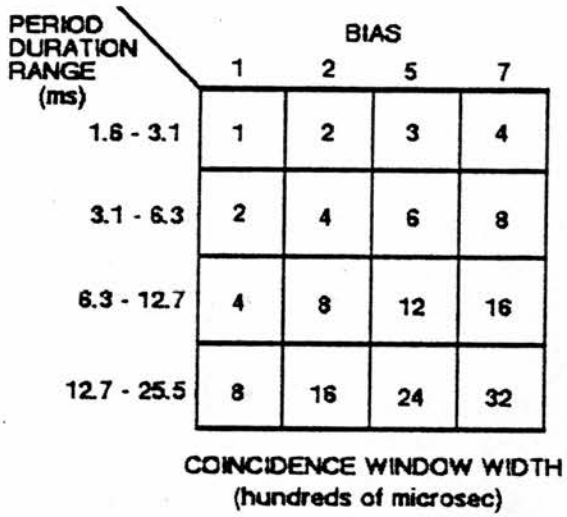


Figure 3.5 The thresholding scheme used during the matching process. The coincidence window widths displayed in the cells of this table are chosen as a function of the duration of the candidate pitch period (the ranges listed down the left of the table) and the matching bias (the numbers listed across the top of the table). (After Gold and Rabiner 1969).

3.5. Following the registration of all possible coincidences for a given primary candidate at a given coincidence window width, a bias factor is subtracted from that particular total of registered coincidences. For example, a bias of 1 is subtracted from the total coincidence count found for a primary candidate on the first and strictest comparisons for coincidences. The biasing factors compensate for the increased numbers of coincidences which occur for a primary candidate as the threshold accuracy is decreased. The importance of the biasing factors in determining the voiced/unvoiced classification of the most likely period duration is discussed in the following section. In the course of completing the matching process for deriving the most likely pitch period, a total of 6 (the primary candidates) * 4 (the increasing window widths) * 35 (the remaining estimates for comparison) coincidence measurements must be completed.

The final selection of the most likely period is completed as follows. At each of the 4 tolerance levels, choose the primary candidate amongst the six possibilities which has the highest number of coincidences. Subtract the appropriate biasing factors from each of the four selections. The period candidate with the greatest number of coincidences amongst the 4 unbiased selections is chosen as the most likely pitch period.

Two processing factors should be noted for the matching process. Firstly, the parallel processor discards some pitch period information at the onsets of voicing in speech -- 3 sets of period estimates (rows 1-3 of the matrix) are required prior to the accurate measurement of the most likely period duration, thus resulting in the loss of the first 2 period estimates. Secondly,

there is a bias in the matching process towards selecting the period data contained in periodicity function m1 (the positive peak maximum amplitude impulse train) since it is the primary candidate to be examined first during the comparisons for coincidences. This second point will be discussed further in the Section 3.2.3 below on pitch period markers.

SECTION 3.1.9 -- VOICED/UNVOICED DECISION

The matching process described in the previous section is also used to determine if the selected most likely pitch period represents a voiced segment of speech (Gold 1964). A voiced/unvoiced discrimination is determined from the level of agreement between the primary candidate and the remaining estimates of the matrix as found for the 4 tolerance levels (the level of agreement is the total number of coincidences for the winning candidate). Figure 3.6 (after Gold 1964:1660) presents the general scheme behind the voiced/unvoiced decision logic. Each row of the Figure represents probability densities for the number of coincidences associated with segments of speech which have been visually determined as being voiced or unvoiced in some given experiment. Within each row, two distributions of probabilities are seen, one for voiced speech (the dashed line) and the other for unvoiced speech (the solid line). The number of coincidences for voiced speech tend to be high since many of the pitch estimates contained within the matching matrix represent redundant information about the periodic behavior of the waveform. Coincidence measures derived from unvoiced speech segments tend to be low, demonstrating a lack of redundancy for the noise-like signals. Each row of the

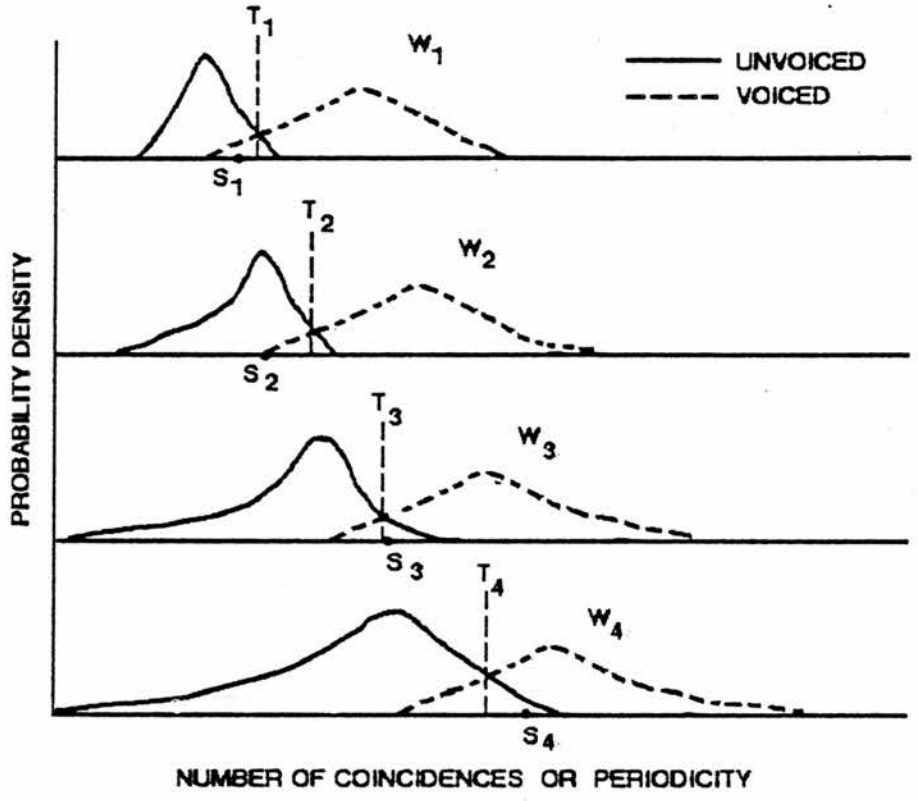


Figure 3.6 The general scheme of the decision logic used by the parallel processor to produce voiced/unvoiced classifications. Each row shows probability distributions of the number of coincidences (or periodicities) associated with voiced and unvoiced speech signals. The number of coincidences for voiced and unvoiced speech increases as the window width increases (represented by the indexed W in each row). (After Gold 1964).

Figure displays the probability distribution for the coincidences measured from the voiced and unvoiced speech as the tolerance levels (i.e. the coincidence window widths W1-W4) have been increased. Note that both voiced and unvoiced distributions move to the right as the number of acceptable coincidences increases for the two types of speech due to the increased tolerance levels. In each row, the point marked by the indexed T is the decision threshold for determining whether a segment is voiced or unvoiced. The decision thresholds are equivalent to the biasing factors discussed in the previous section and were determined by empirical research done by Gold and Rabiner. Given this set of decision thresholds, a new previously unknown set of coincidence measures derived for the 4 tolerance levels (the indexed S in the Figure) may be compared to the decision thresholds (i.e. the biasing factors are subtracted). If the resultant unbiased coincidence count is greater than or equal to zero then the most likely pitch period represents a voiced segment (see, for example, S3 and S4 in Figure 3.6); otherwise it is unvoiced (S1 and S2 in the Figure). As noted in the previous section, the greatest unbiased coincidence count is used when determining the voiced/unvoiced decision.

Three points need to be made in regard to the voiced/unvoiced discrimination logic. Firstly, the present implementation of the parallel processor permits modification of the bias factors/decision thresholds for improving the voiced/unvoiced classification. This capability was included in the design since it is possible that the bias factors established by Gold and Rabiner (1969) may not completely apply to speech samples used in the present study (e.g. due to differences in recording equipment and conditions). The bias

factors may be modified by a value which is added to each of the decision thresholds. In the present study, an additive factor of 4 (i.e. biasing factors of 5, 6, 9 and 11) gives satisfactory results for voiced/unvoiced classification of speech produced by speakers with healthy phonation mechanisms.

Secondly, a most likely pitch period estimate labeled as unvoiced is represented by a zero value upon output from the discrimination process. A series of 3 or more unvoiced decisions is considered a substantial duration of unvoicing which causes all the appropriate basic extractor buffers to be cleared (including the smoothed estimate of period Pav and the 6x6 pitch period matching matrix). This buffer clearance leads to the loss of pitch period data at voicing onsets as noted in the previous section.

Finally, the problems associated with the type of voiced/unvoiced discrimination used by the parallel processor needs to be considered. In this system, periodicity measures are used for determining both the most likely pitch period and the voiced/unvoiced classification. In a comparative study, Rabiner et al. (1976) examined the performance of a number of time domain and spectral domain PDAs including the parallel processor. For voicing discrimination errors, the parallel processor performed as well as the other types of PDA. Most of the detectors used in this study relied on periodicity measures to determine the voiced/unvoiced classification. However, it is perhaps more helpful to separate the two tasks completely, that is, pitch detection and voicing detection. Hess (1983) notes a truism in that pitch can only exist where voicing exists -- this statement is not reversible such that voicing is present only when pitch can be detected. For example, it

may be very difficult to detect periodic components in the voiced speech produced by pathological speakers who use a harsh phonation type. Thus, voicing determination based on "strength of periodicity" (Siegel and Bessey 1982) measures alone may not be completely satisfactory. In addition, each PDA will react to certain signal types in an erroneous manner thus misclassifying voiced speech as unvoiced due to the detector's inability to find periodicity. Therefore, the voiced/unvoiced discriminator used in the parallel processor may be the detector's greatest algorithmic shortcoming. More sophisticated algorithms for voicing classification (see, for example, Rabiner and Sambur 1977; Rabiner, Schmidt and Atal 1977; Siegel and Steiglitz 1976; Siegel and Bessey 1982) would be useful for improving the present system since they incorporate a variety of spectral energy and periodicity measures in rigorous classification schemes.

SECTION 3.2 — ANALYSIS CONDITIONS FOR OBTAINING PITCH DATA

Since the main objective of the perturbation analysis system is the capture of valid cycle-to-cycle perturbation information, a number of analysis conditions linked to the pitch period detection process need to be considered. The general approach behind the implementation of the parallel processor is to apply the system to an interval of speech data, accept the last pitch period within an analysis interval detected by the exponential decay system as the representative period, and then shift the interval forwards in time to include the next pitch period. The analysis conditions of most importance to the system are thus the nature of the analysis interval, the shifting of the interval, and the waveform feature to

be used as pitch period markers.

SECTION 3.2.1 -- ANALYSIS INTERVAL CONDITIONS

Each pitch period estimation is completed on a segment of filtered speech data selected by a rectangular analysis window. The interval within the window is set to accommodate the largest probable pitch period to be produced by a given speaker. At present, the analysis interval is set to a duration of 25 ms for male speakers and 20 ms for female voices. The choice of these durations for the analysis interval determines the absolute lower end of the pitch range which can be detected by the parallel processing PDA. Since three sampled data points are required to define a peak minimum or maximum in the pre-processed speech signal contained within an analysis interval, sampled data points are lost at the beginning and end of the interval. Therefore, the longest possible pitch period permitted into a given interval is slightly smaller than the length of the window. For example, a 25 ms analysis interval can contain a maximum duration pitch period of 24.8 ms (40.3 Hz) at a sampling rate of 10 KHz or 24.9 ms (40.2 Hz) at a 20 KHz sampling rate. For the 20 ms analysis interval, the longest possible pitch period is 19.8 ms (50.5 Hz) at 10 KHz or 19.9 ms (50.3 Hz) at a sampling rate of 20 KHz. Given the rather long durations of the analysis interval, it is normal for more than one pitch period to be present in the window at any one occasion of period detection. The program has been designed to produce an estimate of period for the last complete cycle in the window.

SECTION 3.2.2 -- SHIFTING OF THE ANALYSIS INTERVAL

Cycle-to-cycle data is estimated by shifting the rectangular window along the data in such a way as to try to bring just one new pitch period into the window. A shift of 10 ms (100 samples at 10 KHz sampling rate) would thus be ideal for a steady fundamental frequency of 100 Hz. However, this ideal situation is seldom reached because, in continuous speech, fundamental frequency is naturally moving up and down, both for intonational reasons and for perturbatory reasons. The algorithm is therefore accurate, in the estimation of any two adjacent periods, only within a certain band of fundamental frequencies. The limits of this band are set by the size of the shift factor, basically. If one considers the situation where a new cycle is being brought into the window by one application of the shift factor, then the longest new period that can be accurately detected is one which is no longer than the shift factor itself. If it is longer, then the previous cycle, already estimated once, remains the last complete cycle in the window, and is re-reported. Under-shifts of the interval thus result in over-reporting of pitch periods. Conversely, the shortest new period that can be accurately detected is one which is, at a minimum, greater than half the shift factor itself. If it is half the duration or shorter, then (assuming that the next cycle has the same period or less) the algorithm effectively jumps a cycle and reports the next one as the last in the window. Over-shifting of the interval therefore results in under-reporting of pitch periods. Thus, an octave band of accurate potential F0 estimation is provided by a given shift factor -- this band demonstrating tolerance to increased F0s and intolerance to decreased F0s, relative to the shift factor. This is perhaps less important if one's interest lies in intonation, but it becomes very relevant if the object of

attention is perturbatory behavior, where exact cycle-to-cycle measurement is the goal.

It can be seen that the algorithm retains accuracy of perturbatory tracking only to the extent that the combination of intonational and perturbational movement of F0 remains within a frequency-zone whose limits are determined by the shift factor. It is clearly helpful if a shift factor can be chosen, in the examination of a given voice, that relates in duration to some statistical property of the period durations to be found in that voice, to optimize accurate pitch period estimation. The simplest pitch-adaptive strategy would be to set the shift factor to one value for males, another for females, and another for children, on the basis of general values found in these populations. The next step in tuning the shift factor to allow accuracy of pitch period extraction would be to adjust it to some statistic of the individual speaker's typical performance, for example, the mean, median, or mode F0 of the habitual speech. Finally, one could try to make the shift factor fully pitch-adaptive, using strategies to change the value of the shift factor dynamically, on the basis of predictions about future short-term period behavior reached from examinations of local past short-term history of F0. These three types of pitch-adaptive strategies will be referred to as sex-specific tuning, speaker-specific fixed tuning, and speaker-specific variable tuning.

All three types of approach were used experimentally by Hiller et al. (1983) in comparing the benefits of fixed and variable settings of the shift factor. For each speaker, a preliminary pass through the data was completed, using a sex-specific shift setting

of 10 ms (this setting was used on voice samples of male speakers digitized at 10 KHz). From this, the median F0 was calculated and used to give a fixed shift which was speaker-specific. Alternatively, the sex-specific setting was used as a starting-point for processing the speaker's data by means of a variable shift factor. This variable shift was calculated as follows:

1) An assumption was made that there is an underlying orderliness in the train of pitch periods in speech. In the extreme case this would be represented by an F0 contour which would be a straight line — level, rising, or falling. Within voices that can be considered to be normal and healthy, perturbatory excursions can be anticipated to be infrequent, to be small in extent, and to have a normal distribution for size of excursion.

2) What was needed was some means of projecting the predicted slope of the F0 trend, from knowledge of recent F0 trend behavior. One possibility was to use a moving-average approach to establish the history of recent F0 trend. But means are very vulnerable to the influence of single eccentric values. So it was decided to base the projection of slope of the F0 on recent medians. A running 5-point median was chosen for the study.

The prediction of slope was calculated as follows: let Sn equal the variable shift factor to be evaluated as an optimized attempt to bring in the next pitch period Pn economically and accurately, and Mn equal the median value of the five estimated periods prior to that next period. Sn can be estimated on the basis of the difference between the two most recent median values (Mn -

Mn-1), this difference being a measure of the slope of the F0 trend as estimated at the appropriate delay for the median (i.e., Pn-3). If the difference is equal to zero (i.e. the projected slope is horizontal), then let the next variable shift Sn equal the previous shift factor Sn-1. Otherwise, the next shift is determined from a straight-line approximation from the last median value which includes a factor for the delay, that is, Sn = Mn + 3(Mn - Mn-1).

With this variable shift, inaccuracies will arise only under certain conditions of F0 movement (leaving aside the consideration of perturbations for the moment). These inaccuracies occur at any intonational corner -- i.e., at any point of departure from a straight-line trend. It can be seen that there are limiting values for accurate measurement in these changing contours, beyond which error is inherent.

Figure 3.7 displays two hypothetical pitch period contours, rising and falling, to which the variable shifting logic has been applied. Each contour (the solid line) is plotted as pitch period duration (ordinate) versus the order of the pitch period estimated sequentially in time (abscissa). The first six points of each contour are the six most recently measured periods. Point Pa on the abscissa is the next period (of as yet unknown duration) to be estimated relative to the shift factor produced by the variable shifting algorithm for medians Mn-1 and Mn. It can be seen for each contour that the zone within which accurate estimation of the incoming period can be achieved (the octave band represented by the dotted line at point Pa) has values determined by the local short-term F0 behavior. In the case of the rising pitch period contour (i.e. falling intonation contour), there is tolerance to

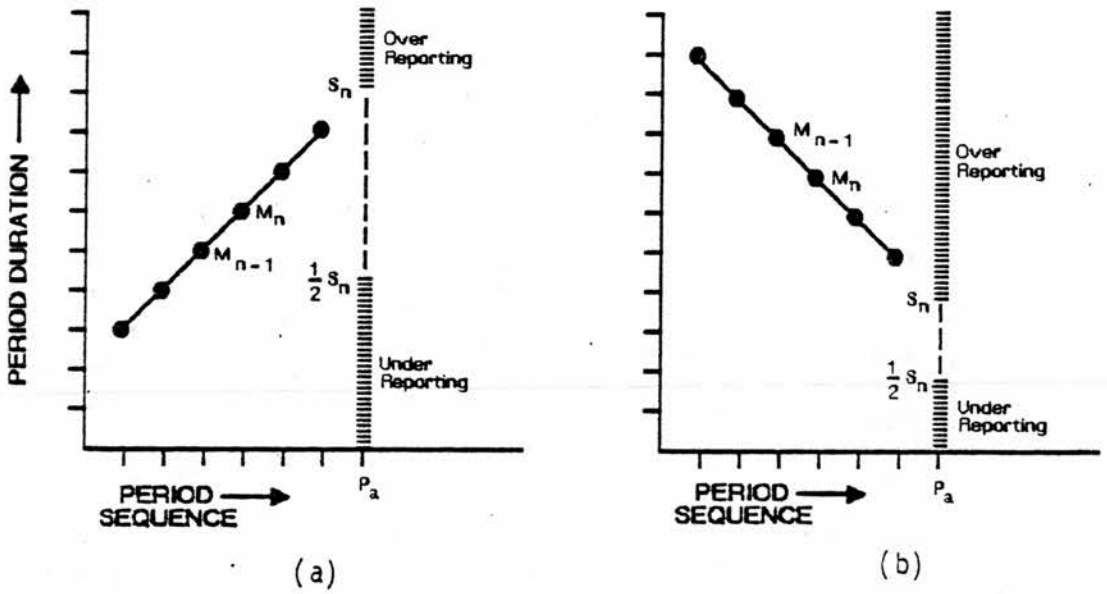


Figure 3.7 Application of the variable shifting algorithm to two pitch period sequences (a — rising, b — falling). The two medians Mn and Mn-1, derived from the six most recently measured periods, are used to predict the next period at point Pa. Note how the octave band (at Pa) for accurate period estimation is located relative to recent F0 behavior.

change-over points (i.e. falling to rising intonation) and no tolerance for rising accelerations of period duration (i.e. increasingly negative intonational slope). For the falling pitch period contour (i.e. rising intonation), there is tolerance for falling accelerations (i.e. increasingly positive intonational slope) and no tolerance for change-over points (i.e. falling to rising periods, rising to falling intonational contour).

Similar constraints operate for perturbed waveforms, and the underlying assumption of orderliness in the data in the form of a straight line tendency becomes progressively invalid with increased severity of cycle-to-cycle perturbatory differences. There are two major problems in severely perturbed waveforms for a variable shifting mechanism of this sort. Firstly, the projection of the predicted slope of F0 can swing wildly, giving values for Sn which take extreme forms and which thus minimize the likelihood of effectively capturing the next true period. Secondly, with contributory adjacent median values differing widely, it is logically possible for negative shifts to occur. In these circumstances, using a variable shift can actually be counterproductive, and can itself contribute artificially to high perturbation values. A partial solution, up to moderate perturbation levels, is to set range-limits. A range-limit of 40 to 240 Hz for male speakers was set. When the extrapolated shift fell outside this limit, the calculation was canceled, and the first-pass speaker-specific value was substituted. At the same time, a flag was set for each occurrence of this out-of-range incident, to keep a measure of how often the range-limits were invoked, and the first-pass speaker-specific shift value substituted.

SECTION 3.2.3 -- PITCH PERIOD MARKERS

In most time domain PDAs, the output of the basic extractor is a series of markers which indicate the boundaries of pitch periods within voiced segments of the speech signal. However, the parallel processor does not operate in this manner. In this case, the output of the basic extractor is an estimate of most likely period duration for a given moment in time -- a matching process being used to select this estimate amongst candidates from 6 parallel channels. It is this matching process which provides the fault-tolerant characteristic of the parallel processing PDA necessary for the analysis of irregularities in speech signals produced with varying degrees of perturbed phonation. If a series of pitch period estimates (or their equivalent F0 values) produced by the parallel processor is an accurate representation of speech waveform behavior then knowledge of the original period markers is not necessary for evaluating the perturbatory movements of the periods.

However, the evaluation of amplitude perturbations (i.e. shimmer) within a given speech waveform will be affected by the loss of the phase relationship between the markers and final estimates of most likely pitch period (Hess (1983) labels this phase relationship between signal and results as marker synchronization). Each amplitude value used to form an A0 contour is based on the peak measure stored in the periodicity function (i.e. m1 -- m6) associated with the parallel channel selected by the matching process. Although the starting points (i.e. period markers) are also to found in the 6 periodicity functions, this information is discarded during the matching process which may choose any one of the six period estimates as the final winner. Therefore, the

resultant associated A0 contour will be composed of amplitude values which vary not only due to actual movements of amplitude in the original signal but also from the switching between parallel channels and the consequent marker asynchronization. Thus, marker asynchronization can lead to less accurate shimmer results. On the other hand, if there is a consistent amount of channel switching during pitch detection then this behavior may reflect irregularities in the input signal which contribute to the resultant A0 contour.

The final computation of the period by matching logic is biased towards the positive peak (i.e. periodicity function m1) when comparisons between the various period measures produce equivalent levels of agreement (e.g. in smooth unperturbed segments of voiced speech). The positive peak parameter is the one most directly related to the impulse behavior of the vibrating vocal folds. Further bias towards the positive peak has been added to the system to accommodate small variations in period measurements. Firstly, it was observed early on that the period durations varied slightly between some of the channels for a given pitch period. The slight variations appear to be the result of actual differences for the various features of the pre-processed waveform, and perhaps of the effects of the digitizing process. In these cases, it was observed that some other pitch period marker (e.g. m4 -- the negative peak minimum) had the highest level of agreement even though positive peak markers were clearly visible and similar in duration. This is the logical consequence of a program which uses past information and redundant features in the matching process to arrive at a final decision. It was decided, after inspection of some typical waveforms, to force the final measure of the period to be from

positive maximum function m1 if the difference in duration between some other chosen feature and the positive measure was minimal. For the time being, the system is set to choose the positive peak marker if there is a difference of less than or equal to 1.5 ms between the two period measures. This minimal difference appears to work well for a majority of cases, as will be discussed below. Differences greater than 1.5 ms are accepted as an indication of perturbed behavior in the waveform and the alternative peak marker duration and amplitude is stored.

Secondly, it was also observed that the speech waveform often appears to be asymmetrical about the zero axis. Hess (1983) noted that the asymmetry of the speech signal is due to recording conditions used to collect the voice sample. Each speaker is recorded close to the microphone and if the first harmonic is strong then asymmetry is found in the resultant voice sample which will favor one of the two polarities. Whether the negative or positive polarity is favored depends on the external recording conditions, for example, the manner in which the recording microphone has been poled. One of the benefits of low-frequency phase compensation is that the asymmetry of the speech signal is preserved. The parallel processor includes a manual switch for flipping the polarities of the input speech waveform in order to bring prominent peaks observed in the negative polarity into the positive polarity. These prominent peaks will then receive the inherent bias of the matching process towards positive peak measures. The decision to flip the signal polarities is determined interactively at the same time as the computation of the silence detection threshold. If the decision is taken to flip the signal polarities then this technique is

applied to the entire voice signal during pitch analysis.

SECTION 3.3 -- PERFORMANCE OF THE MODIFIED PARALLEL PROCESSOR COMPARED TO VISUAL EXAMINATION OF PITCH PERIODS IN SAMPLES OF CONNECTED SPEECH

It was important to evaluate the performance of the pitch period extraction system when applied to data of known characteristics. In particular, one is concerned with the behavior of the system under the two methods of window shifting (fixed and variable) which it was felt would have the greatest effect on accurate perturbation measurement. The following discussion is based on a small study (Hiller et al. 1983) to determine the types of error produced by the automatic system in comparison to visual examinations of speech stimuli.

SECTION 3.3.1 -- THE STUDY

The automatic pitch period detector and visual examinations were applied to the stimulus utterance 'A rainbow is a division of white light into many beautiful colors'. Tape recordings of the utterance were produced by three healthy male adults (RK, JL, SH). In this study, each recorded voice sample was digitized at a sampling rate of 10KHz. In addition, there was no phase compensation of the voice samples during the pre-processing stage. The pre-processing by low-pass filtering was completed by an analog filter (cutoff = 600 Hz) which also prevented aliasing of the input signal during digitization. The parallel processor was applied to the data in two manners: 1) shifting of the analysis window by a fixed speaker-specific shift factor based on the median period

duration derived on a first-pass analysis of the stimulus and 2) variable shifting using a shift factor based on the median shifting logic presented in the section above. The output of the automated system was compared with visual examinations of the low-pass filtered versions of the stimuli using a cursor program on the minicomputer's visual display unit. The results of the comparisons are summarized in Tables 3.1 and 3.2 for the fixed and variable shift conditions for each speaker.

SECTION 3.3.2 -- RESULTS OF THE COMPARISON

Under-reporting/over-shifting; Over-reporting/under-shifting

There is a marginal advantage in these normal voices for the variable shift. In other words, the distribution of F0 values for each speaker falls typically within the accuracy span of the shift-setting of the fixed shift, and making the shift-setting pitch-adaptive brings only a small improvement. It is noteworthy that there is an overall low incidence of pitch period over-reporting for each utterance, given the intolerance of the octave band to F0s deviating towards lower frequencies relative to the local F0 trend. This result suggests that the intonational behavior evidenced in the utterances was mostly free of decelerating changes from the local F0 trends, and that falling intonational contours typically followed more straight-line tendencies.

Over-reporting due to Shimmer Factors in Sudden Low-amplitude Values for Waveform Peaks

Recalling that an exponential decay function is an integral part of the period detection algorithm, when shimmer factors drop the

Subject	RK N=199;CTX=96	JL N=216;CTX=93	SH N=211;CTX=90
Under-reporting/ Over-shifting	5.0% (10)	4.2% (9)	7.1% (15)
Over-reporting/ Under-shifting	2.5% (5)	5.1% (11)	1.4% (3)
Over-reporting/ Low Amplitude	1.5% (3)	1.4% (3)	3.3% (7)
Non-positive Peak Detector	3.5% (7)	3.2% (7)	2.4% (5)
Voice-to-unvoiced Error	0.5% (1)	0.5% (1)	1.4% (3)

Table 3.1 Errors in automatic period detection, using a FIXED shift factor, relative to visual detection, in three healthy male voices. (From Hiller, Laver and Mackenzie 1983:53).

Subject	RK N=199;CTX=96	JL N=216;CTX=93	SH N=211;CTX=90
Under-reporting/ Over-shifting	4.5% (9)	3.7% (8)	3.6% (8)
Over-reporting/ Under-shifting	2.5% (5)	3.2% (7)	2.6% (6)
Over-reporting/ Low Amplitude	1.5% (3)	2.3% (5)	3.1% (7)
Non-positive Peak Detector	3.0% (6)	3.2% (7)	3.1% (7)
Voice-to-unvoiced Error	0.0% (0)	0.5% (1)	1.8% (4)

Table 3.2 Errors in automatic period detection, using a VARIABLE shift factor, relative to visual detection, in three healthy male voices. (From Hiller, Laver and Mackenzie 1983:53).

amplitude of waveform peaks below the exponential threshold, the next true peak is usually beyond the shifted window, and the previously reported cycle is treated as the last complete cycle in the window and re-reported. Values for this type of error were low in both the fixed and the variable shift operations, and the differences were negligible. However, this ability of shimmer factors to contribute to jitter data supports Askenfelt and Hammarberg (1980; 1981) suggestion that one should speak of waveform perturbation, rather than of jitter alone.

Non-positive Pitch Period Marker

Despite the bias towards the positive peak parameter, occasionally some other aspect of the waveform receives the majority vote. The figures are very low in both cases due to the additional forcing logic for small variations between simple pitch period detector durations.

Voiced-to-unvoiced Errors

Occasional low levels of agreement between simple period estimates due to perturbations in the waveform result in an improper unvoiced decision relative to the visual estimation of the waveform. The number of voiced-to-unvoiced errors is very low for the data, and supports the findings of Rabiner et al. (1976).

In an early investigation of the overall accuracy of the parallel processor, Laver et al. (1982) compared the output of the PDA for a sample of speech (2.76 secs in duration) with a visual examination of the same duration. Average fundamental frequencies calculated by PDA and by eye were 131.9 Hz and 134.73 Hz

respectively. A test of replicability was completed, by digitizing and analyzing a 34.4 sec passage from a single recording, on five separate occasions. The range of the average F0 was 1.0 Hz over the five repetitions.

SECTION 3.4 — EFFECTS OF SAMPLING RESOLUTION ON THE TIME DOMAIN ANALYSIS OF FUNDAMENTAL FREQUENCY PERIODS

As described in Section 3.1.2 above, the initial step in the automatic system for measuring waveform perturbations is the conversion of the input voice samples from an analog to digital waveform representation. The digitized data is then input to the pitch detection system operating in the time domain -- peak minima and maxima in the sample waveforms are the basic features of periodicity used by the parallel processor to determine the final pitch period estimates. In the original algorithm described by Gold and Rabiner (1969), a sampling rate of 10 KHz was recommended in order that the parallel processor produces estimates with an accuracy of .1 ms. The 10 KHz sampling rate is suitable for extracting pitch period estimates to be used in vocoder speech systems which produce acceptable synthetic quality. In early research into automatic analysis of waveform perturbations, a sampling rate of 10 KHz was adopted for all voice samples including both female and male voices (Laver et al. 1982; Hiller et al. 1983). In another study (Hiller, Laver and Mackenzie 1984), a question arose as to the intrinsic accuracy of the F0 results based on speech data sampled at 10 KHz. A pitch period detector operating in the time domain is initially limited in its measurement accuracy by the effects of temporal quantization due to the sampling

resolution. For example, a sampling rate of 10 KHz is equivalent to a resolution of .1 ms, thus producing steps of approximately 1 Hz at a 100 Hz F0 (i.e. 99.0, 100.0, 101.0 Hz) and steps of approximately 4 Hz at a F0 of 200 Hz (i.e. 196.0, 200.0, 204.0 Hz). It can be seen that a difference in measuring resolution occurs between typical male and female fundamental frequencies, and this difference affects the measurement accuracy of F0 for these voices. The effects of temporal resolution on perturbation measures were studied by Horii (1979) since it was considered to be the most significant factor in a time domain analysis of speech behavior for pitch data. A normal speaker produced sustained vowel phonations at three F0 levels (150, 210, and 260 Hz) which were digitized at five sampling rates (5, 10, 20, 40, and 80 KHz). Horii found that average jitter magnitude between consecutive pitch periods tended to decrease as sampling rate increased for each F0 level. It was suggested that studies which used poorer resolutions (e.g., 5 KHz) to examine sustained vowel phonations found large jitter values which were inflated by the particular temporal resolution. Jitter ratio measures for a given F0 level (i.e. the average jitter magnitude divided by the average period equivalent to the F0 level) demonstrated constant behavior below a given sampling rate and then increased in value above that sampling rate. This finding suggested that jitter measures of sustained vowel phonations were only at an acceptable level for fundamental frequencies below a maximum frequency level associated with a given sampling frequency. For example, Horii found that a F0 level of 210 Hz was the upper limit for accuracy of jitter measures derived from sustained vowel phonations digitized at a sampling rate of 40 KHz. However, it was concluded that data collected for other combinations of F0 and

sampling rate may still be useful in studies of pathological voice disorders which were expected to produce perturbation behavior which differed from the characteristics of normal phonation. The present analysis scheme must cope with two fundamental frequency characteristics associated with the use of continuous speech for voice samples. Firstly, there are the between-speaker differences in F0 which is particularly notable for the two sexes. Secondly, continuous speech is characterized by within-speaker variation of F0 associated with intonation, word-stress, etc.

Figure 3.8 displays the effects of two sampling rates, 10 and 20 KHz, on the resolution of a number of F0 levels. The resolution measure is labeled Just Noticeable Difference of F0 (JND F0) and measured as a percentage factor to normalize for the differences in F0 levels. The JND F0 measure is calculated as the ratio of the absolute difference between the F0 level and the next possible F0 (this difference is based on the temporal resolution of the given sampling frequency) to the set F0 level:

$$\text{JND F0 in \%} = \frac{|\, F0_n - F0_{n-1} \,|}{F0_n} * 100$$

where F0n is the frequency corresponding to a given period. For example, at a F0 level of 100 Hz, and at a 10 KHz sampling rate, the next possible step up in frequency is approximately 101.0 Hz which is equivalent to a JND F0 of 1%. The JND F0 represents the minimum frequency movement (i.e. perturbation) which can be measured for local small variations in F0 for a given sampling resolution. There are two curves for each of the sampling rates in Figure 3.8 -- a step up in frequency is represented by the upper curve and the lower

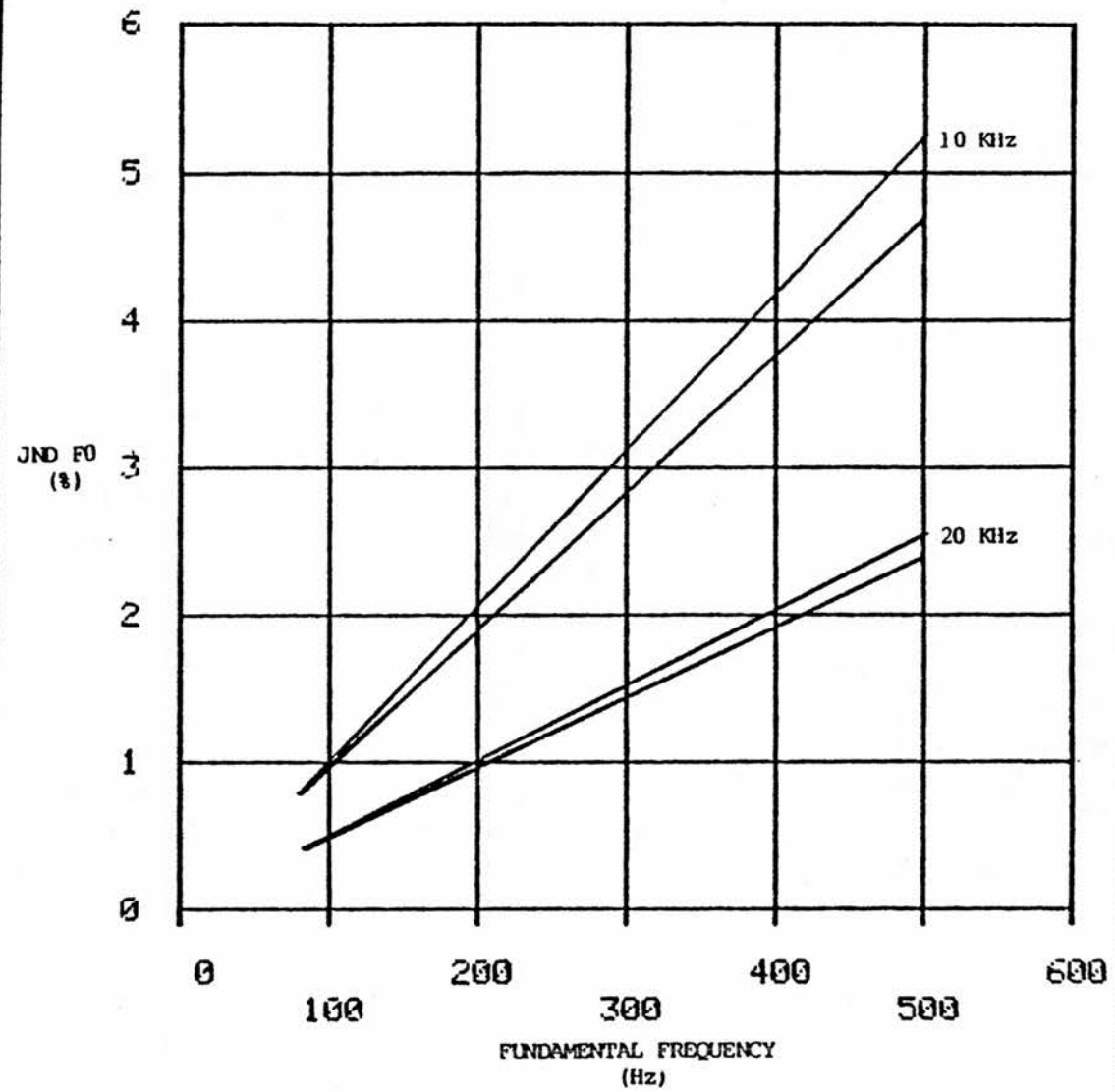


Figure 3.8 Just Noticeable Difference (JND) of F0 in percent plotted as a function of absolute F0 (in units of Hz).

curve represents a step down in F0. It is evident that for both sampling rates increasing F0 level directly controls increasing JND F0. That is, for any given sampling rate, sampling resolution becomes poorer as fundamental frequency is increased. The 'factor of 2' relationship between the two sampling rates is obviously preserved for the resultant JND F0s at each of the F0 levels.

The JND F0 can be used to determine the minimum acceptable error in F0 measurement due to sampling resolution. Hess (1983) reviews psychoacoustic evidence which highlights the acceptable levels of F0 measurement associated with sampling resolution. If the extracted frequency data is to be used as part of a speech synthesis system, then the JND for the audition of F0 changes is the most sensitive indicator of measurement resolution. For example, Holmes (1973) noted that the quantization of pitch due to sampling resolution was audible even for a JND F0 of less than 1%. Hess (1983:85) reported JNDs of F0 which were low for synthetic vowels (0.3 to 0.5% for the male F0 range) compared to the JNDs derived from sinusoidal tones of equivalent frequency. Higher JNDs (4.0 to 5.0%) were noted for real speech stimuli, the greater scores being attributed to the presence of perturbations in the real speech waveforms as compared to the synthetic speech. For experimental tasks other than synthesis, F0 measurements with less accuracy than the JNDs for audition may be acceptable. Hess suggests that the next acceptable level of accuracy could be based on speech production rather than perception, assuming the accuracy of voluntary adjustment of F0 in production is poorer than the perceptual JNDs. A review of a number of perturbation studies leads Hess (1983:85) to conclude that "with respect to accuracy, the ear

thus outperforms the speech production system by far". The magnitude and occurrence of perturbations in normal speech are great enough to exceed the JNDs for F0; the perception of the perturbations being one of phonatory 'roughness' rather than any change of pitch. Finally, the next level of acceptable accuracy of F0 resolution could be based on the linguistic relevance of F0 changes, for instance, in the case of F0 patterns correlated with the perception of stress in English. Hess reports experimental data which suggests that the JND for linguistically-relevant F0 changes (presumably in languages other than tone languages) can be as high as 18 to 25%.

In the present study, interest is focused on the accurate measurement of F0 in order to characterize speakers with normal and dysphonic voices. It is reasonable to select a sampling resolution which will quantize F0 to a JND level typical of voluntary adjustment of F0 production -- this level being greater than the JND for the perception of F0 changes in speech. However, a greater level of resolution accuracy is required if the perturbations in natural speech are to be quantified by acoustic analysis techniques. Thus, a JND F0 of less than or equal to 3% would be a reasonable compromise between measurement accuracy associated with speech production and perception. It can be seen in Figure 3.8 that at the higher frequency levels, a sampling frequency of 20 KHz produces F0 measurement accuracy well below the JND F0 of 3%. Therefore, a sampling rate of 20 KHz would be suitable for quantizing the perturbatory activity of the higher pitches of female speakers and many children.

The use of higher sampling rates does have a number of consequences. Firstly, a much greater amount of digital storage and processing time will be required to analyze long durations of speech data. Secondly, data recorded from speakers with very high F0s may require very high sampling rates in order to detect the presence of perturbations correlated with the early stages of laryngeal pathology. In addition, there is still the problem that any fixed sampling frequency will yield a differential resolution at different fundamental frequency values, both within the performance of a single speaker and between different speakers. Interpolation of the sampled data to increase the apparent sampling rate may be a useful technique for overcoming the overhead associated with increased resolution as well as improving F0 measurement accuracy.

Limiting Pitch Quantization by Interpolation Techniques

Two solutions are available for limiting the inherent quantization of pitch information derived from discrete speech signals. The first solution is to increase the sampling rate when digitizing the time waveform. There are limits to this method since increased sampling rates require increased storage capacity and computation time. In the worst case, a sampling rate of 100 KHz would be required to eliminate audible quantization errors in pitch data extracted from a fundamental frequency of 500 Hz (Hess 1983). The second solution is to use interpolation techniques to increase the resolution of the features of interest in the signal under pitch analysis. The two most-commonly used interpolation procedures are parabolic interpolation and upsampling of the signal. In parabolic interpolation, peak minima and maxima in a signal are treated as

parabolic in shape. Each peak is fitted with a parabola and the new peak location and height is determined from the peak of the parabola. Parabolic interpolation is typically used for pitch detectors working in the frequency domain where the resolution of spectral harmonics is increased by interpolation (see, for example, Martin 1982). Upsampling has been used for pitch detectors working in the time domain (see, for example, Hess 1983; Markel 1972) where the location of features such as peaks and zero-crossings are used as markers of periodicity. Hess (1983) proposed a scheme for upsampling signals to be used in time domain pitch detectors. Firstly, the location of pitch markers are determined in the signal at the original sampling rate. The pitch marker is then upsampled in a narrow region about the marker. The technique of upsampling consists of increasing the sampling rate by the insertion of zeros between existent data points, up to the desired sampling frequency. The upsampled signal is then passed through a low-pass linear phase filter to eliminate unwanted spectral distortions due to the insertion of zeros in the signal (Hess 1983; Rabiner and Schafer 1978). Finally, the output of the filter is searched for the new upsampled location of the pitch marker.

Interpolation of Pitch Markers derived by the Parallel Processor

Parabolic interpolation is used for increasing the resolution of peak minima and maxima derived in the basic extractor stage of the parallel processor. In this case, the three sampled data points which define a given maximum or minimum in the waveform are fitted with a parabola. The resultant peak amplitude and location of the parabola is then used as the basic measures for deriving the 6

functions of periodicity which will be examined for pitch markers by the exponential decay detection system. This is a cross between the two interpolation methods discussed above in that a frequency domain type interpolator is being applied to local features of waveform periodicity to increase the resolution of the peaks. Two reservations exist for this method of interpolating peak features in the time domain waveform. Firstly, no definite statement can be made for the actual increase in resolution provided by parabolic interpolation of time domain features (unlike the upsampling technique in which the number of zeros inserted between sampled data points determines the increase in sampling rate). Davis (1976) and Kasuya, Kobayashi and Kobayashi (1983) also applied parabolic interpolation techniques to time domain waveforms in order to increase the resolution of the resultant pitch period data. Neither study reported the sampling resolutions obtained by using the interpolation procedures. Secondly, parabolic interpolation is assumed to be appropriate to the speech signal which has been considerably smoothed by low-pass filtering prior to pitch marker extraction (i.e. the smoothed peak minima and maxima are parabolic in shape). If this assumption is not correct then parabolic interpolation may give misleading results (though the results will be no worse than peaks derived at the original sampling resolution). In addition, the use of parabolic interpolation requires the use of floating point mathematics to further process the peak features thereby increasing the computational expense of the perturbation measurement system.

SECTION 3.5 — RANGE OF APPLICATIONS FOR THE MODIFIED PARALLEL PROCESSOR

A modified parallel processor for detecting fundamental frequency and amplitude data in connected speech waveforms has been described in this chapter. This time domain PDA is one major component of a system for measuring perturbations of F0 and A0 in samples of connected speech -- this system is to be applied to voice samples produced by healthy and pathological speakers to evaluate its usefulness as a tool for detecting pathological conditions of the larynx. The parallel processing PDA is also suitable for a wide range of other signal processing applications. Hess (1983:3-5) noted 4 general applications in which a PDA is used to measure the voice source in speech signals:

1) Speech Communications — This application area covers many of the speech analysis research topics including transmission and storage, speech synthesis, speaker verification and identification, and speech recognition. Future research on a speech recognition system based on phonetic analysis will use the modified parallel processor for a number of phonetic and linguistic tasks (see next application area).

2) Phonetics and Linguistics -- Fundamental frequency is one of the parameters of interest when analyzing prosodic and phonetic aspects of speech. Hess suggests that the F0 parameter may be influenced by several factors such as the articulatory influence of segmental performance (i.e. microprosody), phonemic tone, word stress, sentence intonation, emotional aspects and speaker characteristics (these factors have been ordered from short- to long-term in nature). The modified parallel processor has been used to examine the tonal and intonational behavior of the Thai language (see

Luksaneeyanawin 1984).

3) Education -- Pitch extraction techniques can be used to provide intonational information for hearing-impaired speakers as well as learners of foreign languages.

4) Medicine and Pathology -- This application is the area of interest for the present study; F0 contours are extracted by the parallel processor for use in a screening system for detecting pathological conditions of the voice. In future research, this type of system may also be applied to the differentiation between types of laryngeal pathology as well as the evaluation of voice rehabilitation through surgical techniques, radiation, chemotherapy and voice therapy. Research is in progress in which the F0 characteristics of stutterers are measured by the modified parallel processor (see Harrington and Hiller 1984 for a preliminary report).

CHAPTER 4

LITERATURE REVIEW OF PERTURBATION STUDIES

CHAPTER 4

LITERATURE REVIEW OF PERTURBATION STUDIES

4.0 INTRODUCTION

A large number of studies (well over 30) have been completed in the investigation of perturbations evidenced in speech waveforms. The following literature review is intended to demonstrate the nature of these studies with particular regard to the experimental techniques used to derive perturbation parameters. This literature review can thus be used as a framework for the detailed discussion of the perturbation measurement algorithms implemented for the present study. Given the large number of studies in the area, it will be seen that few have attempted perturbation analysis of connected speech samples, and even fewer have used a trend line approach as a basis for evaluating frequency and amplitude perturbations.

The following general structure is used to present the development of relevant concepts in the literature of the perturbation analysis field:

1) Period/Frequency Perturbation versus Amplitude Perturbation — the studies reviewed in this section used perturbation measures which are divided into two main types, namely period/frequency and amplitude perturbation parameters. Both types of parameter are measurements of the very short-term cyclic variations in laryngeal vibration displayed in time domain representations of speech signals. A period/frequency perturbation parameter reflects the

degree of regularity of the temporal components of the fundamental frequency of a speech waveform. In all the investigations of period/frequency perturbations, the basic evaluation of temporal regularity uses individual period duration values as extracted from speech signals by manual or automatic techniques. In a few studies, the period duration values are converted to units of Hz for ease of presentation (as is the case in the present study) or units of semitones which are related to the perceptual units used by the human auditory system for pitch perception. This type of parameter will be termed either period perturbation or frequency perturbation depending on the unit of measurement presented in a given study. Jitter is the term often used to describe period or frequency perturbations. The other type of perturbation parameter estimates the regularity of the peak amplitude structure associated with the fundamental frequency of a speech waveform. In all investigations of amplitude perturbations, peak amplitude values are based on the period markers used by pitch extraction techniques to determine the period durations -- the exact operational definition of amplitude varies from study to study (e.g. peak amplitude, peak-to-peak amplitude, etc.). The term shimmer is often used synonymously with amplitude perturbation.

2) The Type of Basic Perturbation Measure — here one is concerned with the basic unit of F0, period or A0 waveform perturbation. A majority of studies use one of two methods for calculating waveform perturbations. One type of perturbation is based on the differences found between adjacent values present in F0, period or A0 contours -- this basic unit of perturbation measurement is labeled as the Cycle-to-Cycle technique. Given a sequence of similar values of F0

or A0 extracted from a speech signal, low difference values are found between adjacent samples which result in a low measure of cycle-to-cycle perturbation (and vice versa for sequences of dissimilar F0 or A0 values). Parameters based on cycle-to-cycle perturbation measures usually consist of a) a measure of rate, that is, the percentage of substantially large differences in a given sequence or b) the average difference between adjacent cycles for a given signal. The average difference value is often divided by the average parameter value for a given sequence in order to normalize the perturbation parameter for the overall level at which the voice sample was produced. It will be seen in a number of studies that the average value of period/frequency perturbation was found to be correlated with the average period/frequency level for a given speech sample. A majority of perturbation studies used a cycle-to-cycle measurement technique for evaluating perturbations of frequency and amplitude.

The other type of unit for measuring perturbations is based on a trend line approach in which a local statistical trend derived from the input contour is used as a baseline from which deviations of the input values can be evaluated for perturbatory behavior. The trend line is used to limit the effects of slow-moving changes of F0 or A0 contours in the measurement of short-term movements of perturbation. In a majority of studies which used a trend line approach, a linear filter consisting of a running-average was the favored method for producing the trend line. Parameters based on trend line perturbation measures usually consist of the average deviation of the input values from their associated smoothed values -- this parameter is also often normalized for overall mean level of

frequency or amplitude. Specialized measures of waveform perturbation to be discussed in this section are for the most part based on cycle-to-cycle or trend line perturbation analysis.

3) The Nature of the Voice Sample -- The original input voice sample from which waveform perturbations are measured usually consists of either sustained vowel productions or samples of connected speech. In a majority of studies, sustained vowel productions were the preferred stimuli. Sustained vowel data, it has been argued, should contain variations due to involuntary laryngeal vibration and be free of variations due to voluntary speech patterns (Hollien, Michel and Doherty 1973). In addition, this type of voice sample requires less sophisticated extraction techniques to derive period and amplitude values -- a sustained vowel phonation can be assumed to be voiced with a repetitive temporal structure for each cycle of vibration (except, of course, in the more extreme instances of irregular phonation). However, voice samples consisting of connected speech have also been investigated since this type of stimulus is considered a more natural use of phonation. That is, a given connected speech sample displays perturbatory characteristics associated with gross changes of phonatory behavior (e.g. at voicing onsets and offsets) as well as finer movements of period and amplitude produced during vowel-like portions of the connected speech. In the case of pathological phonation, the production of connected speech may tax the laryngeal mechanism in a manner which makes the pathology more evident (as seen by increased values of perturbation). The extraction of period and amplitude values from connected speech is more difficult as compared to sustained vowel phonations since the detector must examine a variety of signal

structures produced by interactive effects of the dynamic movements of the speech articulators.

The type of signal analyzed for perturbation measures is also noted for each study reported in this literature review. In general, F0 and A0 perturbation measures have been calculated from airborne signals recorded by microphone or laryngeal signals picked-up at the throat by a contact microphone or miniature accelerometer. The laryngeal signals were often used for perturbation analysis because of their simplified temporal structures which were relatively free of contamination from supralaryngeal resonance effects. The input values to either cycle-to-cycle or trend line perturbation analysis techniques have been derived by manual (i.e. visual) detection of pitch data or by automatic computer programs. Semi-automatic pitch detection techniques also exist in which visual examination of a speech signal is aided by automatic analysis methods (e.g. the use of computer visual display unit with waveform and cursor display programs).

Investigations of waveform perturbations in the human voice output have covered a number of issues. A majority of studies were concerned with the use of perturbation measures to describe the phonatory behavior of healthy and pathological speakers — the main aim of these studies was to discriminate between these 2 groups of speakers. The remaining perturbation studies examined the effects of aging as well as articulatory events (such as tongue height in vowel production) on phonatory efficiency.

SECTION 4.1 — PERIOD/FREQUENCY PERTURBATION (JITTER) STUDIES

In this section, a number of studies are reported in which the main purpose was to describe and quantify perturbations of period or frequency evidenced in speech signals. The parameters derived by period/frequency perturbation analysis are often labeled jitter measures. The discussion of the period/frequency perturbation studies is subdivided according to the experimental considerations described in the introduction to this literature review (see Section 4 above).

SECTION 4.1.1 -- CYCLE-TO-CYCLE PERIOD/FREQUENCY PERTURBATION ANALYSIS BASED ON DATA EXTRACTED FROM SAMPLES OF CONNECTED SPEECH

This section is a review of studies in which perturbation parameters are based on cycle-to-cycle analysis of period duration or fundamental frequency. In most of the studies reported here, a manual technique based on visual examination of waveforms was used to extract period values which were the input to the perturbation analysis. Each study reported here used voice samples consisting of connected speech utterances as the stimuli for perturbation analysis. There are only six studies reviewed in this section, which represents most of the research into perturbation analysis of connected speech. There is a historical aspect to this section as the studies presented here are some of the earliest to be completed in the area -- the techniques used in these studies for perturbation measurement are germinal ones on cycle-to-cycle analysis from which subsequent studies are basic variations.

Lieberman (1961) completed an early study into the nature of the rapid fluctuations of periodic behavior evidenced in acoustic speech waveforms. In this study, the types, frequencies, and magnitudes of the period fluctuations were described as well as the effects of various speech events on periodic behavior. Connected speech samples were produced by healthy male speakers as the stimuli for the study. Each speaker produced the same neutral sentence with eight different emotional modes (bored statement, confidential communication, question expressing disbelief or doubt, fear, happiness, objective question, objective statement, pompous statement). Though not specifically stated, it would appear that each voice sample was tape recorded via a microphone. A group of naive listeners chose those sentences which were perceived as the most identifiable with each emotion. Each selected sentence was reproduced as an oscillogram and visually examined for periods durations. Some general observations were made by Lieberman about the cycle-to-cycle behavior of the sentences. Three patterns of cycle-to-cycle behavior were observed including: 1) undulating -- small increases and decreases of adjacent period duration, 2) steady -- no fluctuations of period duration, and 3) smooth transitions -- either smooth increases or decreases of adjacent period durations. Lieberman determined the incidence of these three cycle-to-cycle patterns by calculating the magnitude of the differences between adjacent periods as seen in sequences of 3 adjacent periods. A large proportion of the cycle-to-cycle behavior of the entire set of sentences was accounted for by undulations and smooth transitions of period (86% of all cases examined). The magnitude of the difference between the durations of adjacent periods was greater than 0.6 ms (1.7 Hz) 20% of the time and greater than 1.0 ms (1 Hz) 15% of the

time. Lieberman compared the values of cycle-to-cycle differences against the values of cycle-to-every-other-cycle differences to determine if the non-steady-state sequences of periods were random movements. Results for this comparison demonstrated low correlations for cycle-to-cycle differences, but cycle-to-every-other-cycle differences demonstrated high correlations. Lieberman concluded that some sort of ordered behavior was present in the periods of the test sentences. Two other fluctuation patterns were noted by Lieberman for the sentences. Firstly, it appeared that as the duration of the periods increased, the magnitude of the differences between adjacent periods also increased. Secondly, large magnitude differences were found at the onsets and offsets of voiced segments as well as during rapid spectral shifts in the speech. Lieberman concluded that these various F0 fluctuation characteristics should be considered when attempting to improve the quality of synthesized speech.

Lieberman (1963) provided further information regarding the small, rapid fluctuations of periodicity evidenced in connected speech. In this study, speech samples were recorded from healthy and pathological speakers which were used to determine the relationship between acoustic waveform perturbations and the vibration of the vocal folds. The resulting information was used in an attempt to differentiate between the healthy and pathological speakers as well as to determine the causes of differing vibratory behaviors. Samples of connected speech were elicited from healthy speakers and speakers with a variety of disorders of the larynx (by and large, these pathologies were laryngeal growths such as tumors and polyps). Each voice sample was tape recorded via a microphone.

Oscillographic tracings were produced for the speech data and visually inspected for periods within the voiced regions. The resultant period durations were input to a computer which completed the perturbation analysis by determining the differences in duration between adjacent periods.

Lieberman first determined the physiological origin of very large magnitude differences between adjacent periods which he observed in the speech data of both groups of speakers. Was the origin of large perturbations in the action of the vocal folds or from the motion of the vocal tract during continuous speech? Lieberman completed the following small experiment to answer this question. Firstly, a healthy speaker produced a sentence using an artificial larynx as the phonation source. Examination of the resultant acoustic waveform revealed no perturbations greater than 0.5 ms in magnitude. Secondly, the speaker produced the sentence with normal phonation, but an external load was applied to the speaker's vocal tract. The acoustic waveform evidenced very large perturbations, some exceeding 0.5 ms in magnitude. The final step incorporated both artificial larynx source and external loading of the tract which resulted in only small perturbations of the waveform. Lieberman concluded that most perturbations originated from the motion of the vocal folds and that some large perturbations may have been the result of the coupling of the vocal tract to the glottis during connected speech production.

A comparison was made of perturbation data derived from the voice samples of the healthy and pathological speakers. For both groups, it was noted that the magnitude differences between adjacent periods increased as the duration of the periods increased.

Comparisons of healthy speakers with pathological speakers of equivalent median F0s displayed greater occurrences of larger cycle-to-cycle magnitude differences for the pathological speakers. To evaluate this observation, Lieberman derived a Perturbation Factor (PF) for each speaker which determined the percentage of all magnitude differences between adjacent cycles of greater than 0.5 ms (this magnitude was based on the findings of the experiment which used the artificial larynx and external load). The perturbation factor was plotted against median F0 to normalize for the differing F0 levels of the speakers. It was found that for increasing size of growth on the vocal folds, there was a greater separation of the PF scores of the pathological speakers from the scores of the healthy speakers (this finding only applied in cases where the growth did not interfere with the normal closure of the folds during vibration). However, for very large growths, analysis of the acoustic waveforms for periods was very difficult due to their noisy-like appearance and therefore the perturbation factors were not sensitive to the pathologies. The perturbation factor was found to be sensitive to the locations of the growths in and around the larynx. Comparisons of high-speed photography data of glottal vibration patterns of the two groups demonstrated that the acoustic waveforms recorded at the lips reflected the fluctuations of the vibrations of the vocal folds. Lieberman concluded that this evidence warranted further research into the use of acoustic measures as a screening technique for the detection of pathological larynges.

Smith and Lieberman (1969) presented an investigation into the relationship of period perturbations to certain types of laryngeal pathologies. Voice samples were analyzed for a large group of pathological and healthy speakers though specific data were reported only for a subset of pathological speakers who evidenced carcinoma of the true vocal folds. Voice samples were in the form of four sentences (two statements and two questions) produced by each speaker. Each sentence was tape recorded via a microphone and reproduced in oscillographic form (on black and white film) for visual examination of periodic activity within all voiced segments of the data. Smith and Lieberman reported highly reliable visual analyses of periods from sentence to sentence for each speaker. The period data were entered into a computer for perturbational analysis. The analysis program determined the distribution of perturbations for each sentence and each speaker. A period perturbation was defined as the absolute magnitude difference between adjacent periods. A perturbation factor was also calculated as the percentage of perturbations in a given speech sample greater than the perturbational measurement error for this system (i.e., in this study greater than 0.2 ms). As a normalization procedure, Smith and Lieberman compared the period perturbation factor with the corresponding average period duration of each speaker in order to differentiate the pathological speakers from the healthy speakers. Results of this comparison for a group of speakers diagnosed with carcinoma of the true vocal folds revealed that a substantial majority of these speakers had quite different perturbation factors from those of the healthy speakers. Smith and Lieberman concluded that it seemed possible to differentiate several types of laryngeal disorder by acoustic and computer techniques.

Hecker and Kreul (1971) investigated the F0 characteristics of acoustic data recorded from speakers diagnosed with cancer of the vocal folds. The F0 data were analyzed for possible indicators of laryngeal pathology as well as providing information about the functioning of healthy larynges. Two groups of speakers were used in this study -- one group contained pathological speakers with carcinoma of the true vocal folds and the other group consisted of healthy control speakers who were matched to the pathological speakers for age and median F0. Each voice sample consisted of a sentence taken from a passage read by each speaker. Each sentence was recorded as an oscillogram and visually inspected for several aspects of F0 behavior (the onsets and offsets of phonation were eliminated from the data). The maximum rate of change of F0 for each sentence was used to determine whether the pathological speakers demonstrated the quick changes of F0 evidenced by the control speakers. Results for this measure revealed that the pathological speakers tended towards a lower maximum rate of change when compared to the control speakers. A second measure evaluated the presence of F0 perturbations in consonantal sections of the speech waveform which may have been related to the effects of coupling between the vocal tract and glottis. A weak trend was noted for the production of a greater number of perturbations by the pathological speakers as compared to results of the control speakers. Hecker and Kreul suggested that the vocal tract-glottis interaction was different for the pathological speakers based on this trend.

Two perturbation measures were used to evaluate the F0 fluctuations for all the voiced segments of the speech data without regard to phonetic context. The first perturbation measure was similar to Lieberman's (1963) Perturbation Factor which determined the percentage of all cycle-to-cycle magnitude differences greater than 0.5 ms. The two groups of speakers did not appear to be differentiated by this perturbation factor. Hecker and Kreul introduced the Directional Perturbation Factor (DPF) which evaluated the percentage of cycle-to-cycle differences which evidenced a change of algebraic sign (each evaluation of directional change is naturally dependent on adjacent measures of cycle-to-cycle differences). A significant difference for the DPF was found between the two groups of speakers. Sensitivity to type of laryngeal disorder was suggested by Hecker and Kreul as a possible explanation for the differing discrimination results of the two perturbation factors. Lieberman's PF appeared to be more sensitive to the presence of masses on the vocal folds while the DPF may have been more sensitive to the invasion of the folds by a malignancy. Distributions of the F0s was also calculated for each test sentence. Comparisons of distributions for the two groups demonstrated that the pathological speakers had more restricted ranges of F0. Hecker and Kreul suggested that the restricted ranges were caused by edema associated with laryngeal carcinoma and the resultant suppressed laryngeal vibrations.

All the studies reported so far used manual techniques to extract period data from samples of connected speech. With the recent advent of fast and sophisticated computerized techniques for pitch extraction, more extensive research into this area has been

completed. Two studies of pathological speech completed by Askenfelt and Hammarberg (1980; 1981) used an automatic method of detection and measurement of cycle-to-cycle variations of the speech waveforms produced during connected speech. A brief explanation is given of the automated perturbation measurement system prior to a reporting the results of the two studies.

The input to the automatic system consists of a 40 sec sample of connected speech based on a standard read passage. The acoustic signal to be analyzed is picked up by a contact microphone attached to the throat of the speaker and recorded onto standard tape recording equipment. The recorded signal is played into a hardware pitch detection device which is connected to a general purpose computer. The detector is of the envelope modeling type (see Section 2.1.2 of Chapter 2 above for details) in which major periodic features within the waveform are detected by exponential decay circuits operating in both polarities of the signal. Period durations are calculated from the distances between detected period markers, converted to frequency values in units of Hz and digitized onto the computer for perturbation analysis. The pitch detection system produces a resolution of 1 Hz for each F0 value within the measuring range of 90 to 300 Hz. The simplicity of the detection algorithm has implications for the measurement of perturbations within the F0 contour. The envelope modeling system does not incorporate a circuit for detecting the locations of the peak values in a waveform and therefore period markers are based on instances when the speech signal value intercepts the decay function. This thresholding technique leads to the situation where cycle-to-cycle variations of amplitude are reflected in the resultant F0 contour

values. That is, the pitch detector does not provide an adequate separation of amplitude and frequency characteristics displayed in speech waveforms -- Askenfelt and Hammarberg labeled the perturbation parameters based on the output of this detection system as 'waveform perturbation' parameters which reflect both frequency and amplitude variations. The two studies are included in the frequency perturbation section of this literature review since the pitch extractor is a time domain device which produces F0 values upon output. However, it is recognized that complete separation of amplitude and frequency components has not been achieved in many of the investigations of waveform perturbation which used time domain algorithms. This is also the case, to a lesser degree, for the parallel processing system described (see Section 3.1.7 of Chapter 3 above) which is used for deriving frequency and amplitude contours for perturbational analysis in the present study.

A three-step procedure was completed in order to determine the locations of waveform perturbations within the F0 contour produced by the extractor described above. Firstly, all unvoiced segments of a given F0 contour are eliminated by applying an energy threshold to an intensity signal which was calculated from the input speech signal. Secondly, the F0 contour is scanned to determine regions of waveform perturbations. The scanning technique, described by Askenfelt and Sjölin (1980), uses a measure of the Rate of Change of Fundamental Frequency (RCFF) to eliminate non-perturbed segments of the F0 contour. The RCFF is equivalent to the maximum absolute rate of change of F0 for a given time span normalized by the average F0 of the given time span:

$$RCFF = \frac{\frac{\Delta F_o}{\overline{F_o}}}{w} \cdot 100$$

RCFF is described in units of percentage change per ms. To determine waveform perturbations, the time span (w) was chosen to be equivalent to one period such that absolute differences are calculated on a cycle-to-cycle basis. Only sections of the F0 contour which demonstrate RCFFs of greater than 2% per ms for a minimum section of five consecutive F0 values are evaluated for waveform perturbation parameters.

In the study by Askenfelt and Hammarberg (1980), two waveform perturbation factors were calculated for a given F0 contour. The Perturbation Factor (PF) was calculated as the ratio of the number of F0 values which demonstrated large perturbations (as defined by the RCFF measure) to the total number of F0 values in the pre-processed F0 contour:

$$PF = \frac{N_{wp}}{N_{tot}} \cdot 100$$

The Perturbation Magnitude (PM) was calculated as the average absolute magnitude difference between consecutive F0 values found within the designated waveform perturbation regions:

$$PM = \frac{1}{N_{wp}} \sum^{N_{wp}} \frac{|f_{n+1} - f_n|}{f_n} \cdot 100$$

The product of PF and PM, called the Waveform Perturbation (WP) measure, was used in comparison with perceptual data since Askenfelt and Hammarberg felt that auditory impressions of voice abnormality would depend on the combination of the two perturbation factors. A

small experiment was completed in which the perturbation measures were estimated for four pathological speakers. These speakers were also rated for degree of abnormality by a group of listeners. The resultant perceptual estimates were compared to the WP product for each speaker. The results suggested a positive relationship between degree of abnormality and the WP measure. Post-therapy evaluations of the four speakers also demonstrated that the perceptual and acoustic measures had decreased.

Askenfelt and Hammarberg (1981) presented an extension of their waveform perturbation research by the introduction of an additional set of waveform perturbation measures. As in the previous study, no distinction was made between perturbations of period duration and amplitude, both being grouped as waveform perturbations. Seven measures of waveform periodicity were used in this study. The first three measures, Perturbation Factor, Perturbation Magnitude, and Perturbation Product (PP -- originally labeled WP) were the same as reported in the original work of Askenfelt and Hammarberg (1980). These measures provided information regarding the frequency and magnitude of waveform perturbations evidenced in the pre-processed F0 contour. The remaining four perturbation measures required all F0s present in the F0 contour. The fourth measure was Hecker and Kreuls' (1971) Directional Perturbation Factor which determined the frequency of changes in algebraic sign between consecutive F0 values. The remaining three perturbation parameters were distributional measures of the relative Frequency Differences (DF0) which occurred between consecutive F0 values. Delta F0 Distribution Width (DF0W) was the dispersion of DF0s in terms of the standard deviation of the data. Delta F0 Zero (DF0Z) was a description of

the peakedness of the DF0 distribution around the minimum difference of zero. Delta F0 Peak (DF0P) was the ratio of DF0Z to DF0W for the F0 contour. DF0P incorporated the behavior of DF0Z and DF0W since it was expected that the value of the parameter would be high for peaked and narrow distributions derived from healthy speakers' phonations versus a low value associated with less peaked and wider distributions of DF0.

The seven perturbation measures were tested on a group of pathological speakers. The usefulness of the seven measures was to be evaluated by pre- and post-therapy results for the group of speakers. Selection of the speakers was based on perceptual evaluations of pre- and post-therapy voice quality. The criterion for acoustic analysis of a voice was that each voice demonstrated a high degree of perceptual vocal abnormality prior to therapy as well as post-therapeutic improvement of perceptual voice quality. Each speaker was evaluated by a group of listeners for a given set of voice qualities. An overall assessment of degree of abnormality was assigned to each speaker.

The usefulness of each perturbation measure was evaluated for the degree of correlation between the acoustic factor and the perceptual rating of the voice. In addition, the ability to discriminate between pre- and post-therapy conditions of the speakers was evaluated for each perturbation factor. These two parameter evaluations were used to determine the most useful perturbation factors for voice analysis. The PF demonstrated a reasonably high correlation for both men and women as well as the highest discrimination factor for the women and the second highest for men. PM provided the lowest correlation of any factor;

Askenfelt and Hammarberg concluded that PM provided little information about the condition of the voice. Similar results were found for the PP which was expected since it was mainly a product of PF. The DPF revealed the highest correlation of all the measures for male voices. However, DPF demonstrated poor discriminability of pre- and post-therapy voices of both sexes. This finding was due to the narrow range of values of DPF for the group of speakers thus revealing poor information about the status of the voices. DF0W revealed a high correlation with perceptual data as well as a rather high discrimination factor between pre- and post-therapy voices. The DF0Z provided less information than DF0W. DF0P provided information somewhere in between DF0W and DF0Z since it was based on both factors. Results for PF and DF0W suggested to Askenfelt and Hammarberg that the quantity of larger perturbations is what should be emphasized in acoustic perturbation measures. Rank ordering of the results was accomplished by the combination of correlation and discrimination data for each factor. It was found that PF, PP, and DF0W were the best candidates for evaluating the pre- and post-therapy results. However, Askenfelt and Hammarberg noted that PP and PF could be biased by occasional instances of extreme perturbations as well as by choice of F0 detection instrumentation. Conversely, DF0W was a measure of the general distribution of the F0 data and therefore not wholly affected by extremes in perturbation measures. Askenfelt and Hammarberg concluded that DF0W was the best choice for perturbation measures for clinical purposes.

SECTION 4.1.2 -- CYCLE-TO-CYCLE PERIOD/FREQUENCY PERTURBATION ANALYSIS BASED DATA EXTRACTED FROM SUSTAINED VOWEL PHONATIONS

This section is a review of the majority of investigations into the perturbation characteristics displayed in speech signals. In these studies, manual or automatic pitch extraction techniques have been applied to stimuli which consisted of sustained vowel phonations in order to provide the input values to perturbation analysis. The large number of studies found in this area can be explained by the less sophisticated pitch extraction procedures required to analyze sustained phonations — for this type of voice sample, one can assume the presence of voicing and a similar temporal structure from cycle-to-cycle of vibration (except in the most severe examples of perturbed phonation). Typical perturbation parameters calculated in these include measures of rate such as the Lieberman (1963) Perturbation Factor or the Hecker and Kreul (1971) DPF as well as the average magnitude differences found between consecutive cycles in a given phonation. The average cycle-to-cycle perturbation parameter is often divided by the average period duration in order to normalize for the overall level of vibration. Most of the articles to be reviewed here presented results for the phonatory efficiency displayed in speech samples produced by healthy and pathological speakers.

Iwata and von Leden (1970) examined the period perturbation behavior displayed in the phonations of healthy and pathological speakers. The group of pathological speakers included cases of chronic laryngitis, benign tumors, malignant tumors, unilateral paralysis of the vocal folds (some with teflon injections), and myasthenia laryngis. Each speaker produced a sustained vowel phonation which was tape recorded via a contact microphone attached to the throat of the speaker. Each recorded signal was reproduced

in oscillographic form on a visicorder and visually examined for period durations. A perturbation was defined as the difference between adjacent periods and this measure was calculated for thirty consecutive periods extracted from each stimulus. Results for each subgroup of speakers were presented in terms of the distribution of the measured perturbations. The distribution of perturbations for the group of healthy speakers confirmed Lieberman's (1963) finding that very small cycle-to-cycle fluctuations were present in the voice samples of healthy speakers -- these period fluctuations were distributed in a symmetrical curve (the perturbation values ranged from -0.6 to +0.5 ms). A slight sex difference was noted for the healthy group in which female voices demonstrated a greater number of smaller perturbations than the males but with an overall wider range. The chronic laryngitis group of speakers demonstrated a slightly wider range of perturbations as compared to the healthy group as well as significantly greater numbers of the larger perturbations. Several disorders were represented by the benign tumor group (including nodules, hematomas, leukoplakia, and papilloma) which demonstrated similar results to the chronic laryngitis speakers. Iwata and von Leden noted a trend in which extent of tumor was associated with range and magnitude of perturbations. That is, the greater the extension of the benign tumor, the larger the range and magnitude of the perturbations. The malignant tumor group demonstrated the widest and most irregular distribution with greater numbers of large perturbations; this group's distribution approximated the behavior of the papilloma case of the benign tumor group. The results for the laryngeal paralysis group were split into two subgroups. Firstly, the paralysis cases which did not receive teflon injection therapy demonstrated an

increased range of perturbations biased towards the larger magnitude perturbations (which were also greater in number than those of the healthy speakers). The speakers who had received injection therapy demonstrated a significantly narrower distribution compared to the non-injected speakers. The injected speakers also had fewer large magnitude perturbations as compared to the group of healthy speakers. It was concluded that healthy speakers produced small cycle-to-cycle fluctuations of period duration in a normally-distributed manner. Period perturbations were significantly larger and more irregular for the laryngeal pathologies. Significant differences were noted between groups of pathologies for various perturbation magnitudes and these differences may be useful for differentiating between voice pathologies. It was also suggested that pre- and post-therapy perturbation measures may be useful in determining success of treatment.

Hollien et al.(1973) reported on a perturbation analysis system which incorporated an improved method for the visual extraction of period durations from oscillographic recordings of voice samples. Correlational analysis of the known F0 characteristics of synthetic stimuli with the results of visual inspections of the synthetic data demonstrated good validity and reliability for the measurement procedure. The measurement system was then applied to voice samples produced by a group of healthy male speakers. Each speaker produced a sustained vowel phonation at four F0 levels including 100, 141, 200, and 282 Hz. Three measures of F0 were calculated including average F0, average jitter and a jitter factor. Each F0 parameter was based on 50 consecutive cycles measured from each vowel

phonation by the visual extraction method. The average jitter parameter is the mean magnitude difference found between the 50 adjacent cycles (converted to units of Hz). Comparisons of the average F0 parameter to the average jitter suggested that the degree of frequency perturbation increased as a function of mean F0 in a manner similar to Weber's law. A Jitter Factor (JF) was developed to normalize for this relationship as follows:

$$JF = \frac{\text{MEAN JITTER}}{\text{MEAN F0}} \cdot 100$$

Slight differences were still found between the jitter factors for the various mean F0 productions. Hollien et al. suggested that the positive relationship between increasing mean F0 and increasing JF was not supportive of Weber's law. It was concluded that JFs of from 0.5 to 1.0 may be expected for sustained vowel phonations produced by healthy male speakers.

Kasprzyk and Gilbert (1975) reported on perturbation behavior associated with differing vowel height. Other researchers (see, for example, Sherman and Linke 1952 and Emanuel and Sansone 1969) found differing auditory perceptions of harshness as well as levels of spectral noise associated with vowels produced with differing tongue height. In particular, high tongue height was associated with low spectral noise and low degrees of perceived harshness while increased spectral noise and perceived harshness were associated with low tongue heights. Kasprzyk and Gilbert measured the amount of periodicity as a function of tongue height for a group of healthy speakers. Each speaker produced five sustained vowel phonations (/i,u,^,ae,a/) which were recorded as oscillograms for visual

inspection of cycle-to-cycle durations. Lieberman's (1963) Perturbation Factor was applied to the data to determine the percentage of magnitude differences between adjacent cycles which were greater than 0.5 ms. Results demonstrated no significant differences between the Perturbation Factors of the various sustained vowel phonations. Kaspryzk and Gilbert suggested that differences in perceived harshness for vowels might have been due to spectral factors such as peak power or intensity rather than period perturbations of the speech waveform.

Kitajima, Tanabe, and Isshiki (1975) investigated the frequency perturbation characteristics associated with healthy and pathological voice productions. In the study, two groups of speakers were evaluated including a group of speakers who evidenced laryngeal cancer and a group of healthy speakers. A sustained vowel phonation was tape recorded from each speaker via a contact microphone attached to the throat of the speaker. Each stimulus was digitized on to a computer and automatically analyzed for period durations (the sampling rate was set to 12.315 KHz and 9-bit quantization). The period detection program was not described in detail though it appears to use peaks as period markers within each voice sample. The pitch detector was used to extract 100 periods from each stimulus -- the period measures were transformed into frequency and then into units of semitones. Kitajima et al. considered the semitone scale to be appropriate since the human auditory system may resolve frequencies in a logarithmic manner. For each sustained phonation, a perturbation measure was calculated as the average absolute magnitude difference between adjacent periods as follows:

$$\overline{\Delta F} = \frac{\sum_{i=1}^{N-i} |F_i - F_{i+1}|}{N-1}$$

where $\overline{\Delta F}$ is the magnitude of perturbation, Fi represents the fundamental frequency in units of semitones and N is the number of periods measured. The results for the sustained phonations demonstrated a limited range of perturbation measures for the healthy speakers while the pathological speakers produced perturbation scores outside the range of the healthy group's scores. Kitajima et al. noted alternating levels of short and long periods for some of the phonations produced by the pathological speakers which they associated with severe hoarseness. It should be noted that Kitajima et al. also analyzed connected speech utterances for frequency perturbation information. The results of these analyses are reported below in Section 4.1.4 on automated trend line analysis of connected speech.

Smith, Weinberg, Lawrence, and Horii (1978) investigated the relationship of period perturbations and perceived roughness measures derived from voice samples produced by a group of male esophageal speakers. Each esophageal speaker produced a sustained vowel phonation which was tape recorded via a contact microphone attached to the throat. For each voice sample, a one sec segment of the data was reproduced as an oscillographic tracing which was visually examined for cycle-to-cycle variations in period duration. Several measures were calculated based on the period durations including mean F0, mean vocal jitter, standard deviation of the vocal jitter, jitter ratio, and the percentage of the total vowel duration identified as periodic. Mean vocal jitter (in units of ms) was defined as the average difference found between consectutive

pairs of periods displayed in each vowel stimulus. Jitter Ratio (JR) was used to normalize the mean vocal jitter of each phonation sample by the average period of that stimulus as follows:

$$JR = \frac{\overline{X}_j}{\overline{X}_p} \cdot 1000$$

where $\overline{X}_j$ is the mean vocal jitter in ms and $\overline{X}_p$ is the mean period in ms. A group of listeners rated the roughness of the phonations by a pair-comparison technique. Acoustically, the vowel tokens were characterized as mostly periodic in behavior with average F0s of less than 90 Hz. Smith et al. found the mean vocal jitter and jitter ratios of the esophageally-produced phonations to be substantially greater in value than jitter measures derived from phonations produced by healthy and pathological speakers as reported in other studies (see, for example, Hollien et al. 1973). The substantially greater jitter measures were expected since the esophageal speakers used unusual vibratory structures to produce the vowels. Good reliability was demonstrated for the listeners' roughness ratings of the sustained vowel phonations. Non-significant results were found for correlations between the listeners' judgments and the various F0 measures. The results of this study contradicted the findings of Wendahl (1966a) and Coleman (1969) who found good correlations between acoustic perturbations and roughness ratings. Smith et al. suggested that the contradiction resulted from the use of unusual phonation stimuli in this study versus well-controlled synthetic stimuli created by Wendahl (1966a) and Coleman (1969). It was concluded that roughness perceptions were comprised of a number of auditory cues and therefore jitter measures would be only partially related to

roughness. Wendahl's (1969) finding that the magnitude of jitter increased as the period duration increased was supported by this study.

Horii (1979) reported on a number of experimental considerations which are of importance when evaluating period perturbations evidenced in sustained vowel phonations. The first part of the study examined the relationship between the magnitude of perturbation and the median F0 displayed by a sustained vowel phonation. A group of healthy male speakers produced sustained vowel phonations at eleven F0 levels ranging from 98 to 298 Hz in one tone steps. Each vowel stimulus was tape recorded via a microphone and digitized on a computer for automatic extraction of perturbation data (the sampling rate was set to 40 KHz with a 16-bit quantization). Periods were extracted from each vowel phonation by a peak picking computer program as described by Horii (1975). Perturbation parameters were based on the absolute magnitude differences between consecutive periods detected in each vowel. The average magnitude difference and the jitter ratio (i.e., the average jitter divided by the average period of the phonation) were calculated for each stimulus. The group mean for the average jitter magnitude demonstrated a tendency for average jitter to decrease as the level of the median F0 increased. The group mean for jitter ratio was found to increase slightly as a function of increasing F0. Horii's findings for the degree of jitter evidenced in sustained vowel phonations produced by healthy speakers supported the results of Hollien et al. (1973). To determine the correlation between non-sequential and sequential F0 measures, Horii compared the distributions of the F0 periods (i.e. the standard deviation) to

the jitter ratio extracted from each vowel stimulus. A positive correlation was demonstrated between F0 distribution and jitter ratio for all the speakers though there were considerable individual variations. Horii concluded that these 2 parameters represented independent or semi-independent characteristics of vocal fold vibration as found in sustained vowel phonations. Some aspects of the distributional behavior of jitter measures were also considered. Most studies which used a measure of average jitter magnitude have disregarded the algebraic sign associated with the calculated differences between consecutive periods. Horii felt that the result of this technique was to skew the data in a positive direction since algebraically signed magnitude differences tended towards a normal distribution around a mean of zero magnitude difference. It appeared that the median jitter magnitude would be a more proper measure of perturbational behavior. Horii noted that the average jitter magnitude for unsigned data was actually determined by the standard deviation of the signed data. Hence mean jitter magnitude of the unsigned data should be interpreted as the magnitude of the dispersion of the signed data. The effects of temporal resolution on perturbation measures were also studied by Horii since it was considered to be the most significant factor in a time domain analysis of speech behavior. The results of this investigation are reported in Chapter 3 above in the discussion of the effects of sampling resolution upon period detection in the time domain.

Horii (1980) reported on period perturbation data extracted from the sustained vowel phonations produced by healthy male adult speakers. Each speaker produced three sustained vowel phonations including /i,a,u/ which were tape recorded via a microphone. Each

vowel stimulus was digitized (sampling rate equal to 40 KHz with 16-bit quantization) and automatically examined by computer for period data. A three sec segment from the middle of each vowel stimulus was analyzed for periods by a peak-picking program as described in Horii (1975). Several parameters were measured for each phonation including average F0 (in units of Hz), standard deviation of F0 (in units of semitones), mean jitter in units of ms, and jitter in percent. The mean jitter was defined as follows:

$$\text{MEAN JITTER} = \frac{1}{N-1} \cdot \sum_{i=1}^{N-1} |P_i - P_{i+1}|$$

where the Pi represent adjacent periods in ms and N is the number of consecutive cycles analyzed. Jitter in percent was normalized for an individual speaker's mean F0 by determining the ratio of mean jitter to mean period in ms times 100. Group mean jitter measures demonstrated significant differences between the three vowels. A significant low-correlation was found between jitter and standard deviation of F0. Horii suggested that some correlation was expected between jitter and F0 distribution since both were measures of F0 periodicity. Further details of this study are reported in Section 4.2.1 below on amplitude perturbation analysis in which Horii reported on the relationship between jitter and shimmer measures.

Murray and Doherty (1980) measured acoustic F0 characteristics from the phonations of healthy and pathological speakers in an attempt to differentiate between the two groups. The pathological group consisted of adult males who evidenced laryngeal cancer of the vocal folds and the control group was comprised of healthy adult males. Each speaker produced a sustained vowel phonation as well as

a read passage which were tape recorded via a microphone. One sentence was selected from the read passage to determine the mean F0 for each speaker. Two perturbation measures were calculated for each sustained vowel phonation. The visual method of Hollien et al. (1973) was used to derive period durations from each voice sample. For each phonation, the Jitter Factor of Hollien et al. was calculated which normalized the mean jitter by the mean F0 where the mean jitter is defined as the average difference between adjacent cycles (converted to units of Hz). Hecker and Kreul's (1971) Directional Perturbation Factor was also used to measure the changes in algebraic sign for the magnitude differences between adjacent cycles. The group averages for mean F0 extracted from the sustained vowel phonations and the sentences were slightly higher for the healthy speakers as compared to the pathological speakers. The group of healthy speakers also revealed significantly lower jitter and DPF values than the pathological group. Discriminant analysis was applied to the parameters in order to group the speakers. The age of the speaker, the mean and standard deviation of F0 extracted from the read passage, mean F0 derived from the sustained vowel phonation, the Jitter Factor, and the Direction Perturbation Factor were the parameters included in the discriminant analysis. It was found that the DPF was the strongest differentiator of the two groups closely followed by the Jitter Factor. This result supports the findings of Hecker and Kreul (1971) for the DPF parameter. The mean and standard deviation of F0 derived from the sentence also proved to be useful in differentiating between the two groups of speakers. Murray and Doherty suggested that this finding was due to the presence of a mass in the vocal folds which lowered F0 and induced high variability. An a posteriori classification of the two

types of speakers (i.e. healthy and pathological) correctly classified 9 out of 10 speakers.

Kempster and Kistler (1983) noted that the significant results of the Murray and Doherty (1980) discriminant analysis seemed improbable in light of the small sample size (5 speakers per group) and large variability of the F0 parameters. Kempster and Kistler re-analyzed the data using the same discriminant analysis technique and an additional statistical procedure. The re-analyses demonstrated nonsignificant results for all the measures. It was suggested that the sample size was too small for parametric analysis and that the discriminant analysis was unable to differentiate between the 2 groups of speakers. In a response to these findings, Murray and Doherty (1983) completed their own re-analysis of the data which also resulted in nonsignificant differences between the 2 groups for discriminant analysis. The disagreement between the original and new findings was attributed to either errors in the original statistical algorithm or a freak computational error during the processing of the data. Murray and Doherty emphasized that the perturbation measures appeared to be different for the 2 groups of speakers despite the nonsignificant discriminant analysis. Therefore, the original study was claimed to support the use of perturbation measures in the study of laryngeal pathology.

Sorenson, Horii, and Leonard (1980) reported the results of a study in which period perturbation measures were extracted from speakers who had been treated with laryngeal topical anesthesia. The purpose of the study was to demonstrate that the

> 'disruption or reduction of laryngeal tactile feedback disrupts intricate frequency control mechanisms and results in deviations from normal laryngeal behaviors'

(1980:274).

Perturbation parameters were measured from data produced by speakers under conditions of anesthesia and no anesthesia -- excessive jitter measures have been associated with abnormal conditions of the phonatory mechanism and therefore should be sensitive enough to reveal the effects of topical anesthesia. A group of healthy speakers produced sustained vowel phonations at eleven different F0 levels ranging from 98 to 298 Hz in one tone steps under anesthetized and non-anesthetized conditions. Each phonation was tape recorded via a microphone. Successful anesthetization was determined by the absence of the cough reflex. It was believed that the topical anesthesia effected the infra- and supra-glottal mucosa as well as the upper tracheal mucosa. Each vowel stimulus was digitized on to a computer (the sampling rate was set to 40 KHz with 16-bit quantization) and automatically analyzed for F0 information. A peak-picking program described in Horii (1975) was used to derive period durations from the middle segments of each phonation. A number of parameters ~~were~~ was calculated for each stimulus including mean F0, standard deviation of F0, mean jitter, and jitter ratio. Mean jitter (in units of usec) was calculated as the mean absolute magnitude difference between adjacent periods of each sustained vowel phonation. Jitter ratio was the normalized version of jitter in which mean jitter was divided by mean period (in ms) of the stimulus multipled by 1000. It was found, for each F0 level, that the group average for mean jitter was greater for the anesthetized condition as compared to the non-anesthetized condition. The group average for all F0 levels revealed that mean jitter produced under the anesthetized condition was almost twice as large as jitter produced under the non-anesthetized condition. Similar findings

were demonstrated for jitter ratio measures. For either speaking condition, individual results demonstrated significantly greater jitter at the higher F0 levels compared to the lower F0 levels. Further, the greatest significant differences between speaking conditions occurred at the higher F0 levels. Comparisons of individual speakers' F0 distributions revealed significantly greater standard deviations for the anesthetized versus the non-anesthetized conditions. A low correlation was found between F0 distribution and jitter ratio. Sorenson et al. concluded that F0 distribution and jitter ratio measures provided useful information about the effects of topical anesthesia on phonation. It was also concluded that deprivation of sensory feedback of the laryngeal mechanism by topical anesthesia could be measured in the perturbation behavior of sustained vowel phonations. However, caution was advised in interpreting the results since the actual extent of sensory deprivation could not be measured for the speakers. In addition, experimental artifacts such as saliva accumulation on the vocal folds should also be considered for this type of study.

Normative data for F0 characteristics associated with the aging vocal mechanism were reported by Wilcox and Horii (1980). Normative data for older speakers is required in order to distinguish between changes in voice quality due to the aging process and changes related to laryngeal pathology. Data derived from older speakers is important since the occurrence of laryngeal carcinoma is greatest at an advanced age. Two groups of speakers were used in this study -- a younger group of 20 male speakers (aged 18.4 to 25.8 years, mean = 23.3) and an older group of 20 male speakers (aged 60.5 to 80.0 years, mean = 69.8). No history of laryngeal pathology was reported

by any of the speakers and hearing levels were appropriate to each speaker's age. Stimuli consisted of 3 sustained vowel phonations (/i/,/a/,/u/) produced by each speaker which was tape recorded via a microphone. A real-time period tracking program, based on a peak picking method in the time domain (Horii 1975) extracted period durations from the stimuli which were sampled at 40 KHz. The average, median, and standard deviation of F0 was measured for each stimulus. Period perturbations were measured as the jitter ratio of the average cycle-to-cycle difference in ms to the average pitch period of each vowel. The jitter ratio is normalized for differing F0 levels by the inclusion of the average period parameter. The results of the F0 and jitter analysis were evaluated statistically by analysis of variance to determine any significant differences in the data. No significant differences were found between the 2 groups of speakers for the 3 F0 parameters. There were significant differences in F0 between the vowels with the low vowel /a/ lower in F0 as compared to the high vowels /i/ and /u/. The jitter ratio was found to be significantly greater for the older group of speakers as compared to the younger group (though there was considerable overlap in values between the 2 groups). In addition, the vowel /u/ had a significantly lower jitter ratio for the 2 groups compared to /a/ and /i/. Wilcox and Horii proposed a number of structural and functional changes in the larynx associated with the aging process which might result in increased jitter values including: 1) reduction in muscle tonus and strength, 2) atrophic thinning of the vocal folds, 3) ossification of laryngeal cartilages, 4) reduced elasticity of laryngeal structures, 5) reduced endocrine function and 6) arteriosclerotic changes in laryngeal blood vessels. It was noted that jitter values obtained for the older speakers were

greater than the younger speakers but lower than jitter ratios reported for pathological speakers. A larger data base for older speakers including females was suggested for future perturbation research.

Benjamin (1981) reported normative data for fundamental frequency measures extracted from the phonations of non-pathological older speakers. The population of older speakers is characterized by anatomical and physiological changes of the larynx associated with the normal aging process. In addition, a high incidence of laryngeal growths and carcinoma is found amongst the older population as compared to younger people. Therefore, normative data is required for the differential specification of F0 characteristics associated with the aging larynx as opposed to pathological changes. Four groups of speakers were used in this study -- 2 groups of younger speakers which consisted of 10 males and 10 females (aged 21-32 years, means = 29.8 and 29.0, respectively) and 2 groups of older speakers including 10 males and 10 females (aged 68-82 years, means = 74.5 and 73.6, respectively). None of the speakers reported a history of laryngeal pathology or smoking. Each speaker's hearing thresholds were typical of his or her age. Speech samples consisted of the fourth and fifth sentences extracted from each speaker's recording of the "Rainbow Passage" as well as a sustained vowel phonation. A hardware device (Tektronix Visipitch) was used to extract F0 data from the sentences in order to calculate F0 distributional and inflectional data for each speaker. These F0 measures included F0 mode, range (undefined), minimum, maximum, the number of upward and downward inflections, and the maximum inflectional change during a 100 ms duration as described by Hecker

and Kreul (1971). A portion of each sustained vowel was recorded in oscillographic form for the calculation of Lieberman's (1963) Perturbation Factor (the percentage of adjacent cycles which differ by .5 ms or greater). In addition, perceptual ratings of pitch were completed for each speaker by a group of listeners. All the F0 distributional and inflectional parameters were found to be significantly different between the younger and older groups of speakers. Modal F0 was lower for the older speakers as compared to the younger speakers. This finding was not consistent with previously reported F0 data, but previous perceptual studies as well as data reported by Benjamin were consistent with lower F0 for older speakers. The older speakers produced greater F0 ranges compared to the ranges of the younger speakers. This finding was also inconsistent with previously reported acoustic data, but Benjamin attributed the inconsistency to differences in type of speech task required of the speakers (e.g. isolated words versus sentences). Greater numbers of inflections as well as greater average maximum inflectional changes were evidenced for the older speakers as compared to the younger speakers. Benjamin noted that measures of inflection would be useful for older speakers since these speakers did not demonstrate reduced parameters whereas Hecker and Kreul (1971) found reduced inflectional measures for pathological speakers.

For the perturbation measures, it was found that the males had significantly greater perturbation factors than the female speakers. The findings also revealed significantly greater perturbation factors for older speakers as compared to younger speakers. Benjamin concluded that the perturbation measure had limited use for

older speakers particularly males since the perturbation factor is known to increase in magnitude as F0 lowers.

Horii (1982) presented further data for period perturbation characteristics associated with sustained vowel phonations. The aim of the study was to determine if there were significant differences in jitter measures between various vowel types. A group of healthy male adults produced a set of vowel phonations which were tape recorded via a miniature accelerometer attached to each speaker's throat. Each sustained vowel phonation was digitized on to a computer and automatically examined for F0 information. The sampling rate was set to 40 KHz with a 16-bit quantization. A peak-picking computer program as described in Horii (1975) extracted period durations from the middle 3 sec interval of each vowel stimulus. F0 parameters extracted from each phonation included mean F0, standard deviation of F0 and jitter ratio. Jitter ratio was calculated as the average absolute magnitude difference between consecutive periods divided by the average period. Analysis of variances for the parameters extracted from the vowels demonstrated no significant differences between vowel types. The non-significant differences between vowel types found in this study do not support the significant differences demonstrated by Horii (1980) for jitter in 3 different vowels. No explanation is given for the differing results though the use of an accelerometer as the signal pick-up device may have been responsible. Horii noted a tendency for higher mean F0s for high vowel as compared to the F0 values derived from low vowels. Additional information on amplitude perturbations as related to varying vowel type was also reported by Horii -- this information is presented in Section 4.2.1 below on cycle-to-cycle

amplitude perturbation analysis.

Ramig and Ringel (1983) examined the relationship between certain age-related changes in body physiology and various fundamental frequency characteristics. Previous studies of the relationship between chronological age and phonation behavior have produced discrepant results -- the discrepancies may be due to physiological differences between speakers within chronological age groups. The subjects in this study consisted of 48 males split into 3 age groups including: young (25-35 years), middle (45-55 years) and old (65-75 years). Within each age group, speakers were divided into 2 types of physical condition termed good and poor. Type of physical condition was accessed on a number of age-related physiological measures including: heart rate, blood pressure, percentage of body fat and vital lung capacity. F0 characteristics were derived from a number of speaking conditions. Mean, standard deviation and range of F0 were based on sustained vowel phonations (/a/,/i/,/u/), a standard read passage and spontaneous speech. A period perturbation parameter was derived from each sustained vowel phonation. The percent jitter was calculated as the sum of the absolute differences between adjacent period durations multipled by the mean fundamental frequency of the phonation (divided by a factor of 10). Two types of vowel phonation were elicited from each speaker including vowel produced with comfortable phonation and vowels produced with maximum duration (this second condition was meant to tax the speaker's phonatory ability). All data were recorded on tape and digitized for computer analysis of F0 (F0 stimuli sampled at 10 KHz and perturbation stimuli sampled at 40 KHz). A peak picking program was used to extract F0 data from the

speech samples (see Sorenson et al. 1980). The results of the physical condition assessment and F0 computations were analyzed by analysis of variance to determine age, physical condition, and age-by-physical condition effects within the data.

Significant differences between speaker groups were revealed for the perturbation measures derived from vowel phonations produced with maximum duration. For the physical condition factor, percent jitter was significantly greater for the speakers in poor physical condition as compared to the speakers in good physical condition. No significant differences were found between speaker groups for the age factor based on period perturbation measures. This finding suggested that age alone as a factor of voice condition had limited value for differentiating between speakers. The addition of physical condition revealed many more differences between speakers in this study. The age-by-physical-condition interaction was not significant for the speakers though the older group revealed the largest differences for this measure. The lack of clear age and age-by-physical condition effects within the data may have been the result of subject selection -- large within age group differences for physical condition may have obscured any real age effects evidenced by the speakers. No significant findings were revealed for percent jitter measures derived from vowels produced at comfortable phonation levels. Therefore, the use of vowels produced with maximum duration supported the notion of taxing the system to elicit possible breakdowns in production.

The vowel types used in the study also produced significant differences in perturbation measures: /a/ was produced with lower jitter as compared to /i/ and /u/. Ramig and Ringel concluded that age-related changes in body physiology are important contributors to certain F0 characteristics.

The automatic detection of perturbation parameters from the speech of healthy and pathological speakers was investigated by Kasuya et al. (1983). Two groups of speakers were used in the study including a group of healthy speakers and a group of speakers who were diagnosed with various types and degrees of laryngeal carcinoma. Each speaker produced a sustained vowel phonation which was tape recorded via a contact microphone attached to the throat. The contact microphone signal was preferred due to its simple temporal structure which is relatively free of influences from the supralaryngeal resonances. Each voice sample was digitized on to a computer at a sampling rate of 10 KHz and 12-bit per sample quantization. Period durations and amplitudes were derived from each phonation by a peak-picking pitch detection algorithm. To aid the detector in locating period markers, a global period estimate was first derived by an autocorrelation PDA which provided a likely starting duration for searching for markers. As each period marker was located by the PDA, parabolic interpolation was applied to the peak in order to increase the resolution of period duration and amplitude estimates. A specialized measurement of period perturbation based on the Burg algorithm of the Maximum Entropy Method (MEM) was then computed for the sequence of period durations derived from a given phonation. Firstly, the difference values between adjacent cycles of data were computed from the sequence of

period durations. Secondly, the MEM was used to calculate the first 7 reflection coefficients from the sequence of period duration difference values. The seven reflection coefficients were then subjected to a statistical discriminant analysis to create a function called the Period Perturbation Index (PPI). In a comparison of PPI values estimated from the voice samples of the two groups of speakers, Kasuya et al. found a reliable separation between speakers who evidenced advanced cases of laryngeal carcinoma and the healthy speakers (though no statistical evidence was reported to support this finding). The PPI values of the healthy speakers overlapped with the period perturbation parameters estimated for speakers with early stage carcinoma. Further results are reported from this study for period perturbation measures derived from sustained vowel phonations using a trend line approach (see Section 4.1.3 below). In addition, the findings from the analysis of amplitude perturbations by cycle-to-cycle and trend line approaches are reported in Sections 4.2.1 and 4.2.2, respectively.

A number of investigations completed by Ludlow and her associates examined the nature of waveform perturbations as found in voice samples produced by pathological speakers. These studies cover a wide range of experimental interests in the perturbation analysis area since 1) both frequency and amplitude perturbation parameters were extracted from speech data, 2) these perturbation parameters were derived by cycle-to-cycle and trend line analysis techniques and 3) the various perturbation measurements were applied to speakers who evidenced a variety of laryngeal pathologies. In keeping with the structure of this literature review, details of specific perturbation analysis techniques will be given in their

appropriate sections. All the studies used the same data collection system and pitch detection algorithm which are explained in brief here. Firstly, the voice sample collected from each speaker was a sustained vowel phonation which was recorded by an FM instrumentation recorder via a microphone. Secondly, each voice sample was notch-filtered to remove the first formant frequency components in order to prevent pitch tracking errors. The output of the filter was input to a pitch detection algorithm which used envelope modeling techniques to find major peaks within the filtered waveforms. A peak-detecting circuit tracked the major peaks in a given waveform which were then extracted by a zero-crossing basic extractor (see Section 2.1.1 of Chapter 2 above for more details of this type of PDA). The resultant period durations and associated peak amplitudes were digitized on to a computer with a 12-bit per sample quantization.

Two cycle-to-cycle frequency perturbation measures were used in the studies. The Mean Frequency Perturbation was computed by summing the absolute differences between adjacent periods and dividing by the total number of periods minus one. To normalize for difference in average period duration of a given voice sample, a Jitter Ratio was computed by dividing the mean frequency perturbation (in units of usec) by the mean period duration (in ms). These parameters were calculated over at least ten blocks of 50 period subaverages for each voice sample.

Ludlow, Coulter and Gentges (1983a) examined the differential sensitivity of the frequency perturbation measures for three different speaker groups. One group consisted of speakers who evidenced neoplastic laryngeal disorders including vocal nodules and

polyps. These neoplastic disorders involve changes in the mass and stiffness of the vocal fold tissues. The two other groups contained speakers with neurological disorders which resulted in similar phonatory dysfunctions such as breathiness and harshness. One neurological group was made up of speakers with Parkinson's Disease -- this is a neuromotor disorder associated with degenerative changes in the neurological system. The other group consisted of speakers with Shy-Drager Syndrome which is a multiple nervous system atrophy leading to denervation of the laryngeal muscles. The two neurological groups were included in the study to determine whether these disorders effect vocal fold tension and position as displayed by abnormal perturbation measures. The speakers in all three groups were matched for age and sex by healthy control speakers. Statistical analyses of frequency perturbation parameters derived from the phonations of all the speakers found the following results. The speakers in the neoplastic group produced phonations with significantly greater mean perturbation and jitter ratio parameters as compared to their matched control speakers. No significant differences were found between the speakers of either neurological group as compared with their match control groups for the two frequency perturbation parameters. Ludlow et al. concluded that these two cycle-to-cycle frequency perturbation parameters were only sensitive to changes in mass and stiffness of the vocal folds which mechanically disturb vibration in a random manner. Further data for this study is given in Section 4.1.3 below on frequency perturbation analysis based on trend line measurements.

Ludlow, Coulter and Gentges (1983b) completed a study to determine if frequency perturbation measures correspond to morphological changes in the vocal folds of speakers undergoing treatment for laryngeal pathology. In this study, nine pathological speakers were examined who evidenced non-malignant lesions of the vocal folds such as polyps, nodules, contact ulcers and edema. Each speaker underwent some type of treatment for the pathology which caused a change in the vocal fold structures as determined by pre- and post-treatment laryngeal examination. The pre- and post-treatment laryngeal examinations resulted in 3 speaker subgroups including 1) 2 speakers classified as normal in laryngeal structure, 2) 5 speakers who displayed improved vocal fold conditions but with some residual pathology and 3) 2 speakers with scarring of the vocal folds due to surgical removal of a given disorder. The mean frequency perturbation parameter was calculated for phonations produced by each speaker at the pre- and post-treatment laryngeal examinations. It was found that the perturbation results were in broad agreement with the speaker subgroupings based on post-treatment laryngeal examination. The two speakers classified as normal in laryngeal structure demonstrated significant reductions in mean frequency perturbation from pre- to post-treatment evaluation. For the speakers who evidenced residual pathology, only one of the five showed significant reduction in mean frequency perturbation. The two speakers with scarring of the vocal folds displayed significant increases in frequency perturbation from pre- to post-treatment. Further details of this study are provided in Section 4.1.3 below on frequency perturbation analysis based on trend line measurement techniques.

Frequency perturbation parameters were used by Ludlow, Naunton and Bassich (1984) in determining whether speakers who evidenced spastic dysphonia would benefit from surgical treatment of the disorder. The usual treatment is the surgical division of the recurrent laryngeal nerve in order to eliminate the spastic dysphonia. A temporary nerve block of the recurrent laryngeal nerve is often used pre-surgically to determine whether a given speaker will benefit from the permanent nerve sectioning. Ludlow et al. investigated speakers with spastic dysphonia to determine if particular perturbation characteristics are displayed by speakers who benefited from laryngeal nerve block and surgical division. The speakers in this study consisted of 4 patients who evidenced spastic dysphonia as well as four age- and sex-matched control speakers. Prior to treatment, mean perturbation measures were derived from voice samples produced by all the speakers. All the cases of spastic dysphonia demonstrated mean frequency perturbation measures which were greater than their matched controls (the differences were significant in 3 out of 4 of the cases). Thus, the frequency perturbation parameter was useful in differentiating between the spastic dysphonic speakers and their matched controls. However, this parameter was found not to be a useful indicator as to which spastic dysphonic speaker would benefit from surgical division of the recurrent laryngeal nerve. Further details of this study are presented for frequency perturbation analysis based on trend line measures (see Section 4.1.3 below) as well in the Sections 4.2.1 and 4.2.2 on the amplitude perturbation analysis, respectively.

Zyski, Bull, McDonald and Johns (1984) completed a study in which they compared a number of waveform perturbation parameters for their effectiveness in differentiating between healthy and pathological voices. Eight parameters of period and amplitude perturbation were accessed for their discriminability, both singly and in combination. In keeping with the structure of this literature review, results for the various types of perturbation parameter will be reported in their appropriate sections. In this section, a brief description of the experimental methodology is given from Zyski et al. Two groups of speakers were evaluated in this study including a group composed of healthy speakers and a group which consisted of speakers diagnosed for a variety of laryngeal pathologies. Each speaker produced a sustained vowel phonation which was transduced by a miniature accelerometer attached to the throat. The output signal of the accelerometer was digitized directly onto a microcomputer at a sampling rate of 100 KHz. An automatic PDA (see McDonald, Zyski, Johns and Bull 1981) was applied to each voice sample to extract the period durations and their associated peak amplitudes.

In this section, results are reported for cycle-to-cycle measures of period perturbation calculated for each voice sample. There were 3 parameters of this type including:

1) Average Pitch Perturbation (APP) which is the average absolute differences between adjacent periods,,

2) Average Percentage Pitch Perturbation (APPP)

$$APPP = \frac{1}{N-1} \cdot \sum_{i=2}^{N} \left(\frac{P_i - P_{i+1}}{P_i} \right) \cdot 100$$

where each difference between adjacent periods Pi and Pi+1 is divided by the period Pi. The total sum of the differences is averaged and multiplied by 100.

3) Lieberman's (1963) Perturbation Factor which is the total percentage of absolute differences between adjacent period greater than 0.5 ms.

A number of statistical findings were revealed for the perturbation parameters. Firstly, there were considerable overlaps of the parametric distributions found for the 2 groups of speakers. Analysis of variance for the APP and APPP parameters demonstrated significant differences between the healthy and pathological speaker groups. The Lieberman measure was not evaluated for group discrimination since a majority of healthy and pathological phonations displayed a value of 0 for this parameter. The three cycle-to-cycle period perturbation parameters were included in a discriminant analysis of all 8 waveform perturbation parameters to determine the rank order of the parameters' discriminability. Of the 8 parameters, four measures contributed significantly to the discrimination of the two groups of speakers including the APPP (the best predictor) and the APP (the fourth ranked predictor). Further details from this study are reported in section on trend line analysis of period perturbations (see Section 4.1.3), cycle-to-cycle analysis of amplitude perturbations (Section 4.2.1) and the trend line analysis of amplitude perturbations (see Section 4.2.2) displayed in sustained vowel phonations.

Horii (1985) presented period and amplitude perturbation measures derived from sustained vowel phonations produced with vocal fry and modal phonation types. It was suggested that since vocal fry phonation has a unique set of physiological, perceptual and acoustic descriptors that jitter and shimmer characteristics should differ from those of modal phonation. A group of healthy male speakers was used to produce modal and vocal fry phonations. Each speaker produced 3 sustained vowel phonations twice, once with modal phonation and the other with vocal fry. The voice samples were tape recorded via a microphone. A 3 sec segment from each stimulus was digitized onto a computer at a sampling rate of 40 KHz and 16-bit per sample quantization. An automatic pitch detection algorithm described in Horii (1975) was used to determine period durations and peak amplitudes based on peak-picking logic. Fundamental frequency parameters were determined for each sample including mean F0 in units of Hz and standard deviation in units of semitones. The group mean F0 values for the modal phonation were mean F0 = 104 Hz and SD = .27 semitones while the group means for vocal fry phonation were mean F0 = 65 Hz and SD = 1.74 semitones.

Two period perturbation parameters were calculated for this study including the mean jitter in ms and the percent jitter (see Horii 1980 in this section for more details). Analyses of variance demonstrated significant differences in group means between the 3 vowels for the modal phonation condition using the 2 perturbation measures. A significant difference was found between the vowel types only for mean jitter in ms in the vocal fry condition. No statistical tests were completed but Horii suggested that vocal fry was characterized by considerably greater jitter than modal

phonation. Further data from this study is presented in Section 4.2.1 below on amplitude perturbation analysis of sustained vowel phonations.

SECTION 4.1.3 -- TREND LINE PERIOD/FREQUENCY PERTURBATION ANALYSIS BASED ON DATA EXTRACTED FROM SUSTAINED VOWEL PHONATIONS

In all the studies discussed so far, period/frequency perturbations evidenced in samples of sustained vowel phonations or connected speech were evaluated on a cycle-to-cycle basis. The investigations reported in this section used a trend line analysis approach to determine period/frequency perturbation parameters. A trend line of period or F0 values is produced by a statistical technique which is intended to limit the effects of slow-moving changes of the values on the measurement of perturbation parameters. A trend line of smoothed period or F0 values is created by applying a filter to the input contour of values -- in most of the studies, the smoothing filter was a linear one consisting of a running-average statistic of a given length of input samples. Perturbations are evaluated as differences (excursions) between individual input values and their associated smoothed values within the trend line. Perturbation parameters based on trend line analysis were usually calculated as the average excursion of the input values from their smoothed trend line values. In this section, results are reported from studies in which trend line perturbation analysis was applied to samples of sustained vowel phonations. As in Section 4.1.2 above, most trend line investigations use sustained vowel phonations as input due to the relative ease of the pitch extraction task. The slow-moving changes

displayed by the sustained vowel phonations are relatively short-term in nature such as vibrato and tremelo. All the studies reported here were concerned with evaluation of the phonatory efficiency of speech produced by healthy and pathological speakers in order to differentiate between the two types.

The seminal article in this area was reported by Koike (1973) in which laryngeal pathology was evaluated by the measurement of period perturbations based on trend line analysis techniques. Acoustic data were elicited from two groups of speakers -- one group was comprised of pathological speakers who evidenced either tumors or laryngeal paralysis (the nature and degree of the pathologies were not specified) and the other group consisted of healthy speakers. Each speaker produced a sustained vowel phonation which was tape recorded via a contact microphone attached to the throat. Each recording was displayed in oscillographic form by a visicorder for visual inspection of periodic activity and the extracted period durations were entered into a computer for perturbation analysis. Period values were extracted from two sections of each vowel phonation including a steady-state portion (32 cycles) and the initiation of the phonation (17 cycles). Period perturbation measures were based on deviations of individual periods away from a smooth trend line of the periods derived by a running-average approach. The smoothed trend line should limit the slow-moving components of F0 in the measurement of the rapid period perturbations. A 3-point running-average was used to calculate the Relative Average Perturbation (RAP) of a given pitch period contour as follows:

$$RAP = \frac{\frac{1}{n-2} \cdot \sum_{i=2}^{n-1} \left| \frac{P_{i-1} + P_i + P_{i+1}}{3} - P_i \right|}{\frac{1}{n} \cdot \sum_{i-1}^{n} P_i}$$

where n is the total number of periods under analysis and Pi represent the period durations. The 3-point running-average is seen as part of the numerator of the RAP equation and this numerator is equivalent to the average absolute perturbation of each one of the periods. The basic measure of perturbation is the absolute difference between a period Pi and its local 3-point average. Preliminary results demonstrated a strong positive correlation between the average absolute magnitude of the perturbations and average period duration of a given phonation. The average period for a given contour is used in the denominator of the RAP equation -- the average period normalizes the perturbation measure for the overall pitch level of a given phonation.

Distributions of the RAP parameter were determined for each group of speakers for the two types of stimuli. The initiation stimuli were subdivided into two types based on the quality of the initiation. Koike described the two qualities as 'soft' and 'breathy'. The distributions of the relative average perturbation measures for the two types of initiation were large with the soft initiation type displaying a wider range than the breathy type of initiation. Both initiation conditions revealed very wide ranges in comparison to the limited range of perturbation values found for the steady-state portions of the stimuli. The group mean value of the RAP parameter was determined for three groups of speakers (i.e. tumor, paralysis and healthy). The group means were found to be significantly different between the three groups for both types of

stimuli. Koike suggested that perturbation data extracted from the initiation of phonation may not be clinically useful due to the wide variability of the measures. Slight overlaps between the distributions of the three groups were also found but Koike concluded that the relative average perturbation measure may still be useful for the screening of laryngeal pathologies.

Davis (1976) investigated an automatic speech analysis system for its usefulness in the early detection of laryngeal pathology. A computer-based scheme was used to extract acoustic parameters from waveforms representative of the glottal source characteristics input to the vocal tract during speech production. The scheme operated in two parts: 1) derivation of a time domain waveform equivalent to the glottal source input by inverse filtering of the acoustic speech signal and 2) extraction of acoustic parameters from the source waveform to be used in the classification of healthy and pathological voices.

In this study, a representative source signal was derived by the technique of inverse filtering (see Chapter 2, Section 2.1.3 for more complete details). A segment of speech is analyzed by linear prediction techniques to determine the filter characteristics of the speech which represent the resonance effects of the vocal tract. These filter characteristics are then applied to the original speech waveform as a digital inverse filter. The resultant filtered signal is termed the residue and is qualitatively correlated with the temporal aspects of vocal fold vibration. The residue signal of a sustained vowel phonation produced by a healthy larynx usually appears as a series of impulses with a minimal amount of noise components between the impulses. Davis chose the residue waveform

since it provided information about periodic glottal activity which is relatively free of the resonance effects of the vocal tract. A peak-picking procedure is applied to the residue signal to determine the locations of the impulses within the waveform. Period durations were then computed for the distances between detected impulses.

A number of acoustic parameters were extracted from the residues of sustained vowel phonations including perturbation measures. Davis presented a general formulation of the running-average approach to the trend line analysis of perturbations displayed in sustained vowel phonations. The Perturbation Quotient (PQ) is calculated as:

$$PQ = \frac{\frac{1}{N-(k-1)} \cdot \sum_{i=1}^{N-(k-1)} \left| \frac{1}{k} \sum_{j=1}^{k} d(i+j-1) - d(i+m) \right|}{\frac{1}{N} \cdot \sum_{i=1}^{N} d(i)}$$

where k is the the length of the running average and m=(k-1)/2 is the relative location of the value to be subtracted from the local running-average. If d(i) is a set of sequential measures of period duration then the Pitch Perturbation Quotient (PPQ) is derived. The PPQ is equivalent to Koike's (1973) RAP measurement of period perturbation where k is equal to 3 and m=1. Davis used a general formulation of the PQ in order to determine the most appropriate length of running-average for the PPQ in a pattern recognition experiment for classifying healthy and pathological speakers. It was found that a running-average of 5 points for the PPQ parameter best reflected perturbation differences between healthy and pathological speakers. It should be noted that a number of other

acoustic parameters were extracted from residue signals for the pattern classification experiment. An Amplitude Perturbation Quotient is discussed below in Section 4.2.2 on amplitude perturbation measurement based on trend line analysis.

This study used two groups of speakers including a group of healthy speakers and a group of speakers who displayed a variety of laryngeal pathologies. Each speaker produced a sustained vowel phonation which was tape recorded via a microphone and this data was digitized on to a computer at a sampling rate of 6.5 KHz. As demonstrated in Section 3.4 above, a sampling rate of 6.5 KHz is considered to be rather low and may not provide enough accuracy for the quantification of frequency perturbations displayed in voice samples produced by healthy and pathological speakers. Feature effectiveness for the acoustic measures was tested for separating the groups based on the best set of analysis conditions. Group mean values for each parameter demonstrated significant differences between the groups of healthy and pathological speakers. Each parameter demonstrated the expected relationship to the groups (e.g., PPQ scores were greater for the pathological speakers compared to the healthy speakers). The rank order of feature effectiveness demonstrated that the PPQ parameter was the best measure for separating the two groups of speakers. Each parameter demonstrated a normal distribution for each group though the perturbation measure was best fitted by a logarithmic distribution curve.

Koike et al. (1977) investigated F0 characteristics of healthy and pathological speakers in order to establish normative data for use in the detection of laryngeal pathology. In addition, Koike et al. were interested in the relationship between the perturbation characteristics displayed in speech signals and the physiological behavior of laryngeal vibration. Two groups of speakers were examined for F0 perturbation data including one group of healthy speakers and a pathological group comprised of a variety of laryngeal pathologies. Each speaker produced a sustained vowel phonation which was tape recorded via a contact microphone attached to the throat. Each recording was digitized on to a computer (sampling rate equal to 20 KHz) and displayed for visual inspection of periodic activity in the waveform. Fifty consecutive cycles of data were extracted from each voice sample as input to the perturbation analysis. Frequency perturbations evidenced in each voice sample were measured by the Frequency Perturbation Quotient (FPQ) as follows:

$$FPQ = \frac{\frac{1}{n-2} \cdot \sum_{i=2}^{n-1} \left| \frac{F_{i-1} + F_i + F_{i+1}}{3} - F_i \right|}{\frac{1}{n} \cdot \sum_{i-1}^{n} F_i}$$

where Fi represents the inverse values of the detected periods and n is the number of cycles analyzed. It can be seen that the FPQ is the frequency version of the RAP developed by Koike (1973). A running 3-point average in the numerator produces a trend line from which individual frequency values can be compared against. The average frequency value of a given voice sample is included in the denominator to normalize the FPQ for a speaker's typical fundamental

frequency level. Resultant group distributions for the FPQ parameter were skewed for both groups of speakers. Koike et al. noted that the skewed distributions might have affected the degree to which the two groups were differentiated and therefore recommended normalization of the distributions by use of the logarithmic values of FPQ. Further discussion of this study requires an explanation of the amplitude perturbation analysis of the voice samples completed by Koike et al. The remainder of this discussion can be found in Section 4.2.2 below on the trend line analysis of amplitude perturbations in sustained vowel phonations.

Deal and Emanuel (1978) investigated the period perturbations evidenced in the phonations of healthy and pathological speakers. Perturbation data was also studied in relation to measures of spectral noise in acoustic signals. Acoustic data were elicited from two groups of speakers. One group was comprised of speakers with laryngeal pathologies which were associated with hoarse voice quality. The other group consisted of healthy speakers who produced phonations with normal as well as simulated rough voice qualities. Each speaker produced a number of sustained vowel phonations (tape recorded via a microphone) which were rated for degree of roughness by a group of listeners. Good inter-judge agreement and intra-judge reliability was demonstrated by the listeners for the vowel data. Deal and Emanuel determined a number of acoustic measures for each stimulus. Sansone and Emanuels' (1970) Spectral Noise Level (SNL) measures were used to estimate the average inter-harmonic noise level in the 100 to 2600 Hz region of the frequency spectrum of each phonation. Sansone and Emanuel demonstrated a strong relationship between the SNL of the 100 to 2600 Hz region and ratings of the

degree of perceived vowel roughness. For perturbation analysis, each voice sample was input to a bandpass filter which was centered at the first harmonic of the signal in order to isolate this component of the waveform. The output of the filter was recorded on a oscillogram for visual extraction of periods from each phonation. Period durations were estimated from a 1 sec steady-state segment of each filtered stimulus. The Period Variability Index (PVI) was calculated for each phonation as follows:

$$PVI = \frac{1}{n} \cdot \sum \left(X_i - \overline{X} \right)^2 \Big/ \overline{X}^2 \cdot 1000$$

where n is the total number of period values, Xi represents the individual period measures and $\overline{X}$ is the mean period duration for a given phonation. It would appear that the PVI contains an element of trend line analysis in that individual perturbation values are based on deviations of period from a long-term mean value of period for the phonation.

The results for the PVI parameter demonstrated significantly greater group mean PVIs for the pathological group and simulated rough phonations of the healthy speakers when compared to the mean PVI derived from normal phonations produced by the healthy speakers. The same general patterns were noted for within vowel class comparisons though not all comparisons produced significant results. Deal and Emanuel suggested that the PVI measure may have limited usefulness when used with sustained vowel phonations due to the non-significant differences for some of the vowels. Clinical usefulness of PVI was also questioned since only a moderate correlation between this parameter and listeners' ratings of

roughness was found for the phonations. The moderate correlation may have been the result of the short durations of the stimuli or that jitter was only partially related to perceived roughness. This partial relationship of jitter to roughness was also reflected by a moderate correlation between SNL and PVI measures. Further data from the study by Deal and Emanuel is presented in the section on amplitude perturbation analysis (see Section 4.2.2 below).

In Section 4.1.2 above on the cycle-to-cycle analysis of period perturbations, results were presented from Kasuya et al. (1983) for the automatic computation of the Period Perturbation Index from sustained vowel phonations produced by healthy and pathological speakers. Further results from that study are given here since a trend line analysis of period perturbations was also applied to the voice samples recorded from the two groups of speakers. The Period Perturbation Quotient (PPQ) based on the 3-point running average technique of Koike (1973) was computed for a series of period duration values derived from a given vowel phonation. The PPQ is the ratio of the mean absolute difference between period values and their local 3-point running average to the mean period duration of the entire sequence of period durations. This trend line analysis technique should determine the rapid variations of periods associated with pathological laryngeal vibration without the effects of the slow and smooth changes in the data. The results of this perturbation analysis demonstrated a reliable separation of the PPQ scores between the healthy speakers and speakers who were diagnosed with advanced degrees of laryngeal carcinoma (though, as in the case of the PPI measure, no statistical evidence was presented to support this finding). An overlap of PPQ

scores was found for the healthy speakers and speakers with laryngeal carcinoma in its early stages of development. These findings for the PPQ are similar and support the results of the perturbation analysis based on the PPI measure. Linear discriminant analysis demonstrated a slightly better separation of the 2 groups of speakers for the PPI as compared to the PPQ parameter though the actual results of the statistical tests were not presented. Further data from this study is presented in Section 4.2.2 on the trend line analysis of amplitude perturbations from sustained vowel phonations.

In Section 4.1.2 above on the cycle-to-cycle analysis of period perturbations, the findings of Ludlow et al. (1983b) were presented for the mean frequency perturbation and jitter ratio parameters computed for sustained vowel phonations produced by healthy and pathological speakers. In this section, further results from Ludlow et al. are reported for the trend line analysis of period perturbations as applied to the vowel phonation stimuli. In the study, the Deviation from Linear Trend (DLT) was computed as a measure of waveform perturbation. This technique is used to remove the effect of slight variations in F0 due to linear trends and intonational patterns which occur during phonation, and are not random in behavior. For a sequence of F0 samples (F0i, F0i+1, F0i+2, ...), the following equation is used to remove the effects of linear trends in F0:

$$DLT_i = \frac{F_{i-2} + F_{i+2}}{2} - F_i$$

The DLT measure determines the differences between alternate values of F0, deleting every other F0 value and therefore should not be

sensitive to instances of diplophonia (i.e. the repetition of pairs of 2 different period durations within the time waveform). The mean DLT value for a series of F0 measures is computed as:

$$\overline{X}\,DLT = \frac{\sum_{i=a}^{b} |DLT_i|}{b - a}$$

The mean DLT for frequency perturbations was used by Ludlow et al. to evaluate the phonatory efficiency of 4 groups of speakers including a healthy control group, a group of speakers evidencing neoplastic disorders (i.e. vocal nodules and polyps), and 2 groups of speakers with neurological disorders (Parkinson's disease and Shy-Drager Syndrome). The object of the study was to determine whether the 3 groups of disordered speakers displayed similar degrees of frequency perturbation in comparison to the control group. The neoplastic group was selected for this study since the disorders should affect the mass and stiffness of the vocal folds. Changes in vocal fold tension are associated with all 3 groups and therefore significant differences should be found between these groups and the control group if the perturbation parameter measures tension. The findings of frequency perturbation as computed by mean DLT were in agreement with the data reported by Ludlow et al. for the cycle-to-cycle measurements of mean frequency perturbation and jitter ratio. That is, the results of the study demonstrated significantly greater DLT measures for the neoplastic group as compared to the control group. No significant differences were found between the 2 neurological groups and the control group. It was concluded that the DLT measure was sensitive only to changes in mass and stiffness of the vocal folds.

In Section 4.1.2 above, data were presented from Zyski et al. (1984) for the cycle-to-cycle analysis of period perturbations evidenced in the phonations of pathological and healthy speakers. In that study, period perturbations were also analyzed by a trend line analysis technique. The Relative Average Pitch Perturbation (RAPP) is taken from the approach of Koike (1973 -- see above) in which a trend line is formed by the application of a 3-point running-average to a set of period durations. The period durations from each vowel phonation were extracted by an automatic PDA. Perturbations are based on differences of individual period values from their associated local trend line value. The average of these differences from the trend line is normalized for pitch level of the phonation by dividing with the average period duration. The statistical analysis of the RAPP values estimated for the two groups of speakers were in agreement with the data provided by the cycle-to-cycle analysis techniques. Firstly, the distribution of the RAPP values estimated for the healthy group of speakers overlapped with the distribution of the pathological speakers. Secondly, analyses of variance revealed a significant difference between the group variances for the RAPP values calculated for the 2 groups of speakers. Thirdly, the RAPP parameter was found to contribute significantly to the discrimination of the 2 groups of speakers in a discriminant analysis based on 8 perturbation parameters -- the RAPP was rank ordered second best behind the Average Percentage Pitch Perturbation measure in this analysis. Further results from this study are presented in sections on cycle-to-cycle and trend line analysis of amplitude perturbations (Sections 4.2.1 and 4.2.2, respectively).

Kane and Wellen (1985) investigated period and amplitude perturbation displayed in the sustained vowel phonations produced by children who evidenced vocal fold nodules. This study was concerned with the lack of perturbation data collected from high pitch speakers particularly in the case of children in whom vocal pathology is common. Each child produced a sustained vowel phonation which was tape recorded via a microphone. A short passage was also read by each child and rated for severity by one expert listener (more listeners would be required to demonstrate the reliability of the severity ratings). Each phonation was digitized onto a computer at a sampling rate of 10 KHz. Period and amplitude perturbation analysis was completed for each voice sample using the system of Davis (1976 -- see this section above). However, the perturbation analysis was applied directly to the speech waveforms since the high FOs of the children resulted in inaccurate measures based on the residue signals of the voice samples. Period perturbations in each phonation were evaluted by the Pitch Perturbation Quotient in which a trend line is formed by a 5-point running-average of period duration values. Correlational analysis was completed for the clinical ratings and the PPQs derived from the voice samples. Clinical judgements of the children's voices ranged from extremely mild to moderate in severity. The PPQ was found to be significantly correlated with the listener's judgements such that rating of severity increased as the period perturbation value increased. Further data for amplitude perturbation analysis of the children's voice samples is presented in Section 4.2.2 below on trend line analysis.

SECTION 4.1.4 -- TREND LINE PERIOD/FREQUENCY PERTURBATION ANALYSIS

BASED ON PERIOD DATA EXTRACTED FROM SAMPLES OF CONNECTED SPEECH

In this section, the applications of trend line perturbation analysis presented in Section 4.1.3 are extended to the evaluation of period/frequency perturbations displayed in samples of connected speech. As in the case of sustained vowel phonations, the trend line analysis of connected speech samples limits the effects of relatively short-term slow-moving changes in fundamental frequency. In addition, the longer-term movements of F0 associated with intonational factors in connected utterances must also be limited from the perturbation analysis by the use of a trend line technique. Only two studies are reported here as very little research has been completed in which trend line perturbation analysis has been applied to samples of connected speech. The investigation of Laver et al. (1982) is the original research from which the present study arises.

In Section 4.1.2 above on the cycle-to-cycle analysis of frequency perturbations, the findings of Kitajima et al. (1975) were presented for the automatic detection of perturbation parameters from sustained vowel phonations produced by healthy and pathological speakers. Further results from that study are presented here for the trend line analysis of frequency perturbations from connected speech samples. The automatic pitch detection system briefly described above was also used by Kitajima et al. to derive frequency values (in units of semitones) as input to the trend line perturbation analysis. For this part of the investigation, the trend line analysis was only completed for phonations produced by the group of healthy speakers. Each healthy speaker uttered a short phrase of connected speech which was composed of all vowel-like sounds. These voice samples were also

tape recorded via a contact microphone attached to the throat of the speaker and digitized on to a computer at a 12.315 KHz sampling rate. The original cycle-to-cycle perturbation measure was altered to eliminate the effects of the slow-moving changes of F0 due to intonation and/or accent observed for the phrases by Kitajima et al. This approach to the trend line analysis of frequency perturbation uses a least squares approximation to fit a smoothed curve to a series of fundamental frequency values. The Magnitude of Perturbation ($\overline{\overline{\Delta F}}$) is computed as follows:

$$\overline{\overline{\Delta F}} = \frac{\sum_{i=3}^{N-2} \left| F_i - \overline{F_i} \right|}{N-4}$$

where $\overline{Fi}$ represents the least square fit for a local series of 5 F0 values:

$$\overline{F_i} = C_{-2} \cdot F_{i-2} + C_{-1} \cdot F_{i-1} + C_0 \cdot F_i + C_1 \cdot F_{i+1} + C_2 \cdot F_{i+2}$$

where Ci represent the coefficients which are calculated by the least squares fitting method. The exact formulation of the least squares fitting technique is not given by Kitajima et al. (1975). The magnitude of perturbation measure is calculated in semitones rather than units of Hz to approximate the logarithmic discrimination behavior of the human auditory system for changes in frequency. Kitajima et al. found a region of .08 to .19 semitones for a group of healthy male and female speakers such that F values greater than .19 semitones could be considered as pathological.

Laver et al. (1982) used a running-average approach to determine a number of perturbation measures in the connected speech utterances of healthy and pathological speakers. A running-average approach was used to cope with intonational movements of F0 and to provide a baseline from which to measure perturbatory excursions of F0. The perturbation measures were based on the concept of excursion of a given F0 measure from a 5-point running-average of the local fundamental frequency values, expressed as a percentage of the mean. The F0 perturbation measures included 1) AVEX -- the average excursion of F0, 2) SDEVEX -- the standard deviation of the range of excursions and 3) RATEX -- the percentage of points in a given sample where the excursion is equal to or greater than 3%. However, the running-average is not considered an adequate statistical technique for coping with the dynamic features of intonation and perturbation displayed in F0 contours extracted from connected speech. In the research to be reported in Chapter 5, a move has been made to the use of a running-median statistic for creating F0 and A0 trend lines.

SECTION 4.2 -- AMPLITUDE PERTURBATION (SHIMMER) STUDIES

In this section, a number of studies are reviewed in which the amplitude structure of speech signals were analyzed for perturbation parameters. The resultant measures produced by the amplitude perturbation analysis are often labeled as shimmer parameters. As in the case of the period/frequency studies (see Section 4.1 above), the various amplitude perturbation studies are subdivided according to the experimental considerations discussed in the introduction to this literature review (see Section 4 above). This section is

somewhat shorter than the previous one since fewer investigations of amplitude perturbations have been completed as compared to the number of period/frequency perturbation studies. In addition, many of the results reported here are from articles which have been already discussed in the various subdivisions of the period/frequency literature review. Therefore, the description of experimental methodology is somewhat limited for these articles but the relationships between findings for period/frequency and amplitude perturbation analyses are discussed. It will be seen in this section that no studies of amplitude perturbations as found in samples of connected speech have been reported in the relevant literature at the time of this writing.

SECTION 4.2.1 -- CYCLE-TO-CYCLE AMPLITUDE PERTURBATION ANALYSIS BASED ON AMPLITUDE DATA EXTRACTED FROM SAMPLES OF SUSTAINED VOWEL PHONATIONS

In this section on amplitude perturbation analysis, some articles are reviewed in which perturbation measures were estimated on a cycle-to-cycle basis. The amplitude data input to the perturbation analyses were derived by the application of automatic or manual pitch extraction techniques to samples of sustained vowel phonations.

von Leden and Koike (1970) reported on a technique for the detection of laryngeal pathology based on amplitude perturbation of the voice. Two groups of speakers were used in this study including a large group of speakers diagnosed for a variety of laryngeal pathologies as well as a group comprised of healthy speakers. Each speaker produced a sustained vowel phonation which was tape recorded

via a contact microphone attached to the throat. von Leden and Koike considered the waveform picked-up by the contact microphone to be physiologically-relevant to laryngeal dysfunction as well as reflecting good signal-to-noise ratios. Each recorded sample was displayed in oscillographic form on a visicorder and visually examined for amplitude peaks within a reasonably consistent portion of the signal. The amplitude measures for each stimulus were entered into a computer which completed an autocorrelational analysis to determine the periodicities of the amplitude modulations. This study is included as an example of cycle-to-cycle perturbation analysis of amplitude since the autocorrelation technique evaluates the correlation between values within the input sequence. In its most correct sense, cycle-to-cycle correlation is only derived for a lag of 1 which evaluates adjacent amplitude measures. The results of each autocorrelation was graphically presented as a correlogram which illustrates the magnitude of the autocorrelation coefficients versus the autocorrelation lags of the coefficients. Peaks within the positive portion of the correlogram reflect positive correlations while peaks within the negative region reflect negative correlations between amplitudes. A total of 15 lags for the autocorrelation procedure was considered significant to demonstrate short-term amplitude modulations.

Four major types of correlogram were found for the speakers in this study and these types were found to generally correlate with various classes of clinical disorder. The four types of correlogram are as follows:

> Type 1: A smooth, decreasing curve from a region of positive correlation to a region of negative correlation. A high positive value was noted for a lag of one which suggested a strong correlation between adjacent amplitudes

of the waveform. A negative peak was seen within the 10 to 15 lag region which may have been associated with vibrato in the waveform. This correlogram was typical of healthy phonation as well as from the phonation of speakers with minor pathological conditions such as mild inflammation or small nodules on the vocal folds.

Type 2: An irregular, decreasing curve from a region of positive correlation to a region of negative correlation. A high positive value was noted for a lag of one, but the remainder of the correlogram was irregularly shaped between two and ten lags. The irregularities of the correlogram suggested irregular short-term perturbations of the waveforms of speakers with benign organic changes which affected the vibratory margins of the vocal folds (e.g., severe inflammation or inflammatory tumors).

Type 3: An irregular curve which demonstrated no significant trend in the periodicities of amplitude modulations (with the possible exception of a correlation at a lag of one). This type of correlogram was observed for cases of incomplete approximations of the vocal folds such as unilateral vocal fold paralysis and benign neoplasms.

Type 4: A correlogram with marked positive and negative correlation peaks which indicated apparent periodic modulations throughout the range of autocorrelation lags. This type of amplitude behavior was found for cases of malignant lesions and large neoplasms.

In the section on the cycle-to-cycle analysis of period perturbations (see Section 4.1.2 above), results are reported from Horii (1980) in which significant differences were found between vowel types (/i,a,u/) for parameters derived from the phonations of healthy speakers. In that study, Horii also calculated a cycle-to-cycle amplitude perturbation parameter for the same data base. The amplitude perturbation parameter was based on the amplitude of the peaks in a given speech signal as found by a peak-picking PDA described in Horii (1975). The shimmer measure is a logarithmic value calculated as follows:

$$\text{SHIMMER IN dB} = \frac{20}{N-1} \cdot \sum_{i=1}^{N-1} \left| \log_{10} \frac{A_i}{A_{i+1}} \right|$$

where Ai represents the peak amplitude of the ith cycle, Ai+1 represents the peak amplitude of the adjacent cycle and N is the number of cycles analyzed. Cycle-to-cycle variations in amplitude are represented by the logarithm of the ratio of the adjacent peak values. Group mean shimmer values demonstrated significant differences in amplitude perturbation between the 3 types of vowel phonation -- this result supports similar findings by Horii for period perturbation parameters. In this case, the /a/ vowel was found to be greater in shimmer than the two other vowels. A significant low correlation was revealed between the period and amplitude perturbation parameters calculated for the vowel phonations. Horii suggested that the low correlation between jitter and shimmer parameters supports the view that both measures are related to similar sets of physical forces which control laryngeal vibration. The low degree of correlation between these parameters meant that the measures did not reflect overly-redundant information.

Data reported from Horii (1982), in the section above on the cycle-to-cycle analysis of period perturbations (see Section 4.1.2) revealed non-significant differences between vowel types for measures of jitter. The cycle-to-cycle shimmer measure described in Horii (1980; see the previous article) was also applied to the vowel phonations recorded by Horii (1982) -- as in the case of the jitter measures, no significant differences were found between vowels based on the shimmer parameter. These non-significant differences between vowel types for both period and amplitude perturbation parameters revealed by Horii (1982) do not agree with the significant differences demonstrated by Horii (1980). The

conflicting results of the two studies were not directly addressed by Horii in the 1982 study. However, overall differences in shimmer values displayed in the 2 studies were largely attributed to the use of a regular microphone (Horii 1980) versus a miniature accelerometer (Horii 1982) to record the signals input to the perturbation analysis. Horii suggested that signals picked-up by the accelerometer attached to the throat are less variable than the intensity variation of airborne voice signals. Therefore, caution should be used when comparing normative data from studies in which differing analysis techniques have been used for measuring perturbations of period and amplitude.

In the above section on the cycle-to-cycle analysis of period perturbations (see Section 4.1.2) data is presented from Kasuya, Kobayashi and Kobayashi (1983) for the automatic computation of the Period Perturbation Index for sustained vowel phonations produced by healthy and pathological speakers. In the same study, a specialized cycle-to-cycle measure of amplitude perturbation was also determined for each voice sample produced by the speakers. The Amplitude Perturbation Index (API) is calculated in a similar manner to the PPI. That is, the Burg algorithm of the Maximum Entropy Method is computed for a series of peak amplitude values detected by automatic analysis of period markers in a given voice sample. This algorithm consists of 2 stages including 1) determination of the difference values between adjacent peak amplitude values and 2) computation of the first 7 reflection coefficients from the newly-formed series of amplitude difference values. Statistical discriminant analysis is then used to develop the function API based on the reflection coefficients of the amplitude differences. Comparisons of the API

parameter calculated from the speakers of the 2 groups demonstrated a reliable separation of the healthy speakers from speakers diagnosed with advanced cases of laryngeal carcinoma (no statistical evidence was presented by Kasuya et al. to support the findings). The distribution of API values estimated from the voice samples produced by speakers with early stage carcinoma overlapped with the API estimates of the healthy speakers. Therefore, the findings for the API parameter were very similar to those of the PPI parameter for the 2 groups of speakers. Kasuya et al. also investigated an amplitude perturbation parameter based on trend line analysis of voice samples. The findings from this investigation are presented in the section below on automatic trend line analysis of amplitude perturbations from sustained vowel phonations (see Section 4.2.2).

In the section above on the cycle-to-cycle analysis of period perturbations (see Section 4.1.2) results were presented from Ludlow et al. (1984) for speakers who underwent treatment for spastic dysphonia. The automated system used in the investigation also determined a cycle-to-cycle amplitude perturbation measure from the sustained vowel phonation produced by each of the 8 speakers. The peak amplitude values of the period markers in each voice sample, as determined by the envelope modeling PDA, were digitized on to a computer as input to the amplitude perturbation measurement algorithm. The Mean Amplitude Perturbation was computed for each phonation as the sum of the absolute differences between consecutive amplitude values divided by the total number of cycles under analysis minus one. This amplitude perturbation parameter is computationally similar to the Mean Frequency Perturbation as described in Ludlow et al (1983a; 1983b; 1984). All 4 cases of

spastic dysphonia demonstrated Mean Amplitude Perturbation parameters which were significantly greater than the values produced by their matched control speakers. However, this cycle-to-cycle measure of amplitude perturbation was not useful for discriminating between the spastic dysphonic speakers who did and did not benefit from surgical treatment of the disorder.

In Sections 4.1.2 and 4.1.3 above, results were presented from Zyski et al. (1984) for period perturbation analysis using cycle-to-cycle and trend line analysis techniques, respectively. The sustained vowel phonations of the healthy and pathological speakers were also evaluated for cycle-to-cycle amplitude perturbations by automatic analysis. Peak amplitude values evidenced in a given stimulus were examined by three measures of perturbation including:

1) Average Amplitude Perturbation (AAP) which is calculated as the average absolute difference between adjacent measurements of peak amplitude,

2) Average Percentage Amplitude Perturbation (APAP)

$$APAP = \frac{1}{N-1} \cdot \sum_{i=2}^{N} \left(\frac{A_i - A_{i+1}}{A_i} \right) \cdot 100$$

where each difference between adjacent peak amplitudes Ai and Ai+1 is divided by the amplitude Ai. An average is calculated for the sum total of the measured differences and multipled by 100.

3) Shimmer (SHIM)

$$SHIM = \sum_{i-1}^{N} 20 \left| \frac{\log \frac{A_{i+1}}{A_i}}{N} \right|$$

where SHIM is the average power difference between consecutive peak amplitudes.

The findings for the 3 shimmer measures are in agreement with the period perturbation data for speaker group differences. Firstly, the distributions of the cycle-to-cycle amplitude perturbation measures revealed overlaps of the perturbation measures derived from the healthy speakers with those of the pathological speakers. Analysis of variance for the APP, APAP and SHIM measures demonstrated significant differences between the groups of speakers for all 3 parameters. In a discriminant analysis of the 2 groups using 8 perturbation parameters, the AAP parameter was found to contribute significantly to the differentiation of healthy and pathological speakers. This parameter was rank ordered third best of the 8 measures -- the AAP was the only amplitude perturbation parameter which demonstrated significance in the discriminant analysis. A trend line analysis technique for amplitude perturbation was also applied by Zyski et al. to the voice samples, the results of which can be found in Section 4.2.2 below.

In Section 4.1.2 above on the cycle-to-cycle perturbation analysis of periods, data was presented from Horii (1985) for the analysis of sustained vowel phonations produced with modal and vocal fry phonation types by healthy male speakers. An amplitude perturbation measure based on cycle-to-cycle analysis was also applied to the phonations. The mean shimmer (see Horii 1980 in this section for details) was derived from peak amplitude values extracted from each voice sample. In an analysis of variance, it was found that the group mean for the shimmer parameter was significantly different between 3 different vowel types produced

with a modal phonation type. However, no significant differences for shimmer were found between vowel types produced with vocal fry. Horii also noted that the mean jitter measure was significantly correlated with the shimmer parameter in both types of phonation while the jitter in percent was only correlated with shimmer in vocal fry. Without statistical proof, Horii noted that the shimmer parameter of the vocal fry was considerably greater than for modal phonation. The finding for both jitter and shimmer suggested that the mechanism of vocal fry phonation is such that stability of frequency and amplitude are considerably limited.

SECTION 4.2.2 -- TREND LINE AMPLITUDE PERTURBATION ANALYSIS BASED ON AMPLITUDE DATA EXTRACTED FROM SAMPLES OF SUSTAINED VOWEL PHONATIONS

In this section, results are reported from a number of studies in which amplitude perturbation parameters were based on trend line analyses of speech signals. The amplitude data input to the perturbation analyses were determined by manual or automatic detection of values from samples of sustained vowel phonations. All the studies in this section were investigations of the phonatory characteristics displayed in voice samples produced by healthy and pathological speakers.

Kitajima and Gould (1976) reported on the measurement of amplitude perturbations present in the phonations of healthy and pathological voices as an indicator for differentiating the two types of voices. Two groups of speakers were used including a group of healthy speakers as well as a group of pathological speakers who evidenced vocal fold polyps of various sizes and locations. Each speaker produced a sustained vowel phonation which was tape recorded

via a microphone. A 360 ms segment of each recorded phonation was low-pass filtered (cutoff frequency equal to 1.5 KHz) and digitized onto a computer at a sampling rate of 20 KHz and 9-bit per sample quantization. Each digitized sample was displayed and visually-inspected for peak amplitudes associated with each period in the signal. A vocal shimmer measure was used to evaluate each phonation and incorporated a 5-point least squares fit to consecutive peak-to-peak measures of amplitude similar to the $\overline{\Delta F}$ measure of Kitajima et al. (1975). The exact formulation of the least square fitting technique is not given in the study. Vocal shimmer is expressed as the mean amplitude difference between consecutive cycles in dB as follows:

$$\text{VOCAL SHIMMER (dB)} = \frac{\sum_{i=1}^{n} \left| 20 \cdot \log \frac{A_{i+1} + \left(A'_i - A'_{i+1}\right)}{A_i} \right|}{n}$$

where n is the total number of amplitude measures for a given utterance and Ai represents the instantaneous peak-to-peak amplitude values. The amplitude difference (i.e. Ai+1/Ai) is normalized by the local trend values A'i and A'i+1 which are derived by a least squares fitting of 5 local values of amplitude. The smooth trend line was determined by a least squares approximation to the peak amplitudes in order to eliminate slow-moving amplitude components in each stimulus. The distribution of shimmer measures were determined for the healthy speakers and a critical region was calculated in which values outside this region were considered not to be normal (upper limit = 0.19 dB shimmer). The distribution of shimmer measures for the pathological group was found to overlap the

distribution of the healthy speakers. Kitajima and Gould suggested that the overlap was due to the variety of polyps evidenced by the pathological speakers especially the smaller polyps which might have had little effect on laryngeal vibration. Comparisons of individual shimmer measures to the critical region of normal scores revealed that a few of the healthy speakers and most of the pathological speakers fell outside the critical region.

In the above section on trend line analysis of period perturbations, results were presented from Davis (1976) for the automatic computation of the Period Perturbation Quotient from sustained vowel phonations produced by healthy and pathological speakers. The PPQ parameter is a measure of the variability of period durations as derived from residue signals produced by inverse filtering of voice samples. Davis found the PPQ to be the most effective feature amongst a number of acoustic parameters for differentiating between the groups of healthy and pathological speakers. The group mean PPQ calculated for the pathological group was significantly greater than the mean PPQ of the group of healthy speakers. In the same study, a trend line analysis technique was also used to derive a measure of amplitude perturbation for each sustained vowel phonation. Given the general formulation of the Perturbation Quotient, if d(i) is a set of sequential measures of peak amplitudes then the Amplitude Perturbation Quotient (APQ) is produced for a given voice sample. The amplitude values used to produce the APQ were based on the peak measures of the period markers detected by the automatic analysis of the residue signal of each speech sample. The resolution of the detected amplitude values was increased by parabolic interpolation of the associated period

marker. For the APQ parameter, Davis determined that a running-average of 5 sequential amplitude values best reflected the amplitude perturbation differences between the 2 groups of speakers. A filter length of 5 was also found for the PPQ trend line analysis of period perturbations. For the pattern recognition experiment, the APQ parameter was ranked as second best behind the PPQ measure for separating pathological speakers from the healthy speakers. For each group of speakers, the APQ measures formed a normal distribution of values which was best fitted by logarithmic distribution curves. The group mean values for the APQ parameter was found to be significantly greater for the pathological group of speakers as compared to the mean APQ calculated for the healthy group. These results for the APQ measure are in agreement with the findings of Davis for the PPQ parameter.

In Section 4.1.3 above on the trend line analysis of period perturbation, data was presented from Koike et al. (1977) for measured derived from sustained vowel phonations produced by healthy and pathological speakers. In that study, trend line analysis for amplitude perturbations was also completed for the vowel stimuli. Peak amplitudes associated with the period in each phonation were visually extracted for the perturbation analysis. Amplitude perturbations displayed in each voice sample were measured by the Amplitude Perturbation Quotient (APQ) as follows:

$$APQ = \frac{\frac{1}{n-11} \cdot \sum_{i=6}^{n-6} \left| \frac{A_{i-5} + A_{i-4} + \ldots + A_i + \ldots + A_{i+5}}{11} - A_i \right|}{\frac{1}{n-1} \cdot \sum_{n-1}^{i=1} A_i}$$

where Ai represent the peak amplitude values and N is the total number of peaks analyzed. An 11-point running-average in the numerator produces a trend line from which individual amplitude values are compared for perturbations. The average amplitude value estimated for a given phonation is used in the denominator to normalize the APQ for a speaker's overall intensity level. As in the case of the FPQ, Koike et al. found the group distributions for the APQ parameter were skewed for the 2 groups of speakers. A more normal distribution of the APQ values were obtained by completing a logarithmic transformation of each APQ value. The degree of separation of the two groups of speakers was estimated by plotting log PPQ versus log APQ values. It was found that the normal speakers occupied a limited space on the plot while the pathological speakers were spread over a much wider region. Koike et al. noted that certain pathological conditions were organized in specific areas on the plot but there were large overlaps between the pathologies. Critical ellipses were fitted to each group distribution in the FPQ/APQ plane based on the mean and distribution of the values. The critical ellipses may be used to determine whether or not the data from a new speaker belongs to the same population.

In the section above on the trend line analysis of period perturbations (see Section 4.1.3), results from Deal and Emanuel (1978) demonstrated a significantly greater mean Period Variability Index for the phonations of a group of pathological speakers as compared to the voice samples produced by a group of healthy speakers. In addition, vowel phonations produced with rough quality by the healthy speakers were significantly greater in period

perturbations as compared to the normal phonations of the same speakers. In the same study, an amplitude perturbation measure was also obtained from each vowel stimulus based on the peak amplitudes of the cycles extracted by visual examination of their oscillographic representations. The Amplitude Variability Index (AVI) is a trend line measure of perturbation calculated as follows:

$$AVI = \log_{10}\left[\frac{1}{n} \cdot \sum \left(X_i - \bar{X}\right)^2 \Big/ \bar{X}^2 \cdot 1000\right]$$

where n is the number of peak amplitude values for a given phonation, Xi represent the individual amplitude measures and $\bar{X}$ represents the mean amplitude value for a given voice sample. The AVI is very similar to the PVI perturbation parameter with the addition of the logarithmic factor. This log factor was included in the computation of the AVI since it produced more linear relationships between amplitude variability of a signal and measures of spectral noise level as well as listeners' ratings of roughness. The results for the AVI parameter were found to be similar to those of the PVI. That is, a significantly greater mean AVI was found for the group of pathological speakers as compared to the mean value evidenced by the healthy group for their normal phonations. In addition, the rough vowel productions of the healthy speakers demonstrated significantly greater measures of amplitude perturbation as compared to their own normal vowel phonations. The measures of spectral noise level and listeners' ratings of roughness displayed moderate positive correlations with the AVI parameter. AVI did appear to be more strongly related to roughness ratings than PVI -- Deal and Emanuel suggested that AVI may be a better index than PVI. Positive, moderate correlations were also found for

comparisons between AVI and PVI as well as the combination of AVI and PVI compared to SNL or roughness ratings. Deal and Emanuel suggested that these results were not surprising since roughness ratings probably reflect total vowel aperiodicity which may not have been entirely accounted for by the AVI and PVI measures. AVI and PVI may have contained redundant information which was not improved by combining the two measures.

In Section 4.2.1 above on the cycle-to-cycle analysis of amplitude perturbations, results were presented from Kasuya et al. (1983) for the automatic computation of the Amplitude Perturbation Index from sustained vowel phonations produced by healthy and pathological speakers. Further results from that study are given here since a trend line analysis of amplitude perturbations was also applied to the voice samples recorded from the two groups of speakers. The Amplitude Perturbation Quotient (APQ) based on the 3-point running-average technique of Koike (1973) was computed for a series of peak amplitude values derived from a given vowel phonation. The APQ is the ratio of the mean absolute difference between amplitude values and their local 3-point running average to the mean amplitude of the entire sequence of amplitude values. The results of this perturbation analysis demonstrated a reliable separation of the APQ scores between the healthy speakers and speakers who were diagnosed with advanced degrees of laryngeal carcinoma (though, as in the case of the API measure, no statistical evidence was presented to support this finding). An overlap of APQ scores was found for the healthy speakers and speakers with laryngeal carcinoma in its early stages of development. These findings for the APQ are similar and support the results of the

perturbation analysis based on the API measure. Linear discriminant analysis demonstrated a slightly better separation of the 2 groups of speakers for the API as compared to the APQ parameter though the actual results of the statistical tests were not presented.

The following summarizes the results of Kasuya et al., 1) cycle-to-cycle and trend line perturbation measures of both period and amplitude produced reliable separation between advanced cases of laryngeal carcinoma and healthy speakers, 2) there were overlaps between the healthy speakers and the early stage cases of laryngeal carcinoma for all the parameters and 3) the cycle-to-cycle measures PPI and API produced slightly better separation of the two groups of speakers as compared to the trend line measures PPQ and APQ.

Period and amplitude perturbation results have been reported from Zyski et al. (1984) in Sections 4.1.2, 4.1.3 and 4.2.1 above. The final data to be discussed from that investigation is for a trend line analysis technique used to extract amplitude perturbation measures from sustained vowel phonations produced by healthy and pathological speakers. This trend line measure is an adaptation of Koike's (1973) RAP as applied to peak amplitude values evidenced in voice samples. The Relative Average Amplitude Perturbation (RAAP) is calculated as follows:

$$RAAP = \frac{\frac{1}{N-2} \cdot \sum_{i=2}^{N-1} \left| \frac{A_{i-1} + A_i + A_{i+1}}{3} \right| - A_i}{\frac{1}{N} \cdot \sum_{i-1}^{N} A_i}$$

where individual measures of amplitude Ai are evaluated for perturbations relative to a local 3-point running-average value of

amplitude. The average of the amplitude differences from the trend line is normalized by the average amplitude value for the voice sample under analysis. The findings for the RAAP parameter are similar to the results reported by Zyski et al. for differentiating pathological from healthy speakers by the other perturbation parameters. Firstly, the distributions of RAAP values for the 2 groups of speakers partially overlapped. Secondly, analysis of variance of the RAAP measure revealed a significant difference between the healthy and pathological group. In a discriminant analysis procedure using 8 measures of perturbation, the RAAP parameter did not contribute significantly to the discrimination of the 2 groups.

In the section above on the trend line analysis of period perturbations (Section 4.1.3), data was reported from Kane and Wellen (1985) for measures derived from phonations produced by children diagnosed with vocal nodules. Significant correlations were found between a listener's ratings of vocal severity and the period perturbation parameter. Amplitude perturbation analysis was also completed for each voice sample using a trend line approach. The Amplitude Perturbation Quotient of Davis (1976 -- see above) was automatically calculated for a sustained vowel phonation produced by each child. The trend line analysis consists of a 5-point running-average from which individual peak amplitude measures can be evaluated (amplitude data was extracted directly from speech waveforms rather than residue signals for these speakers). As was found for the Period Perturbation Quotient, the APQ was found to be significantly correlated with an expert listener's judgement of vocal severity such that increased severity was associated with

increased amplitude perturbation. In addition, the APQ, PPQ and severity rating measures were found to be significantly correlated with each other. Kane and Wellen suggested that the correlation was important for clinical applicability since higher degrees of reliability for objective measures which are sensitive to small changes in phonation.

SUMMARY OF PERTURBATION LITERATURE REVIEW

This section identifies the major concepts, revealed by a review of the perturbation literature, which have had an impact on the development of the perturbation measurement system to be completed in this thesis. These concepts are: perturbation, jitter, shimmer, cycle-to-cycle perturbation, trend line and excursion, phonatory efficiency as evidenced in isolated vowel productions and connected speech, and the use of acoustic analysis of perturbation factors for screening, tracking and diagnostic purposes.

In summary detail, most studies of waveform perturbation can be classified according to two main types of perturbation parameter. A frequency or period perturbation parameter (jitter) quantifies the degree of regularity displayed by the temporal components of the fundamental frequency of the speech signal. Measures of amplitude perturbation (shimmer) evaluate the regularity of the peak amplitude structures associated with the speech waveform's fundamental periodicity. Secondly, two basic units of waveform perturbation have been used to produce frequency and amplitude perturbation parameters. The cycle-to-cycle perturbation describes the relationships between adjacent pulses of vibration as seen in speech

waveforms. In the trend line approach, the basic unit of perturbation is measured as the deviation of F0 or A0 values from equivalent smoothed values produced by a local statistical smoothing algorithm. Thirdly, perturbatory behavior has been observed for two types of speech sample including sustained vowel phonations and samples of connected speech. A majority of studies have measured cycle-to-cycle frequency and amplitude perturbation parameters from samples of sustained vowel phonations. Only a small number of investigations have applied trend line analysis techniques to samples of connected speech in order to derive measures of frequency and amplitude perturbation.

The main purpose in most studies of waveform perturbation has been the use of perturbation parameters as indicators of phonatory efficiency in the vocal behavior of healthy and pathological speakers. Many investigations have attempted broad classification of speakers as healthy or pathological based on frequency and amplitude perturbation measures. A small number of studies have examined specific diagnostic categories of laryngeal pathology in order to characterize each disorder by type and degree of perturbatory behavior. Perturbation parameters have also been used as indicators of the success of therapeutic treatment of laryngeal pathology by medical and/or voice therapy in pre- and post-examinations of phonatory efficiency. In general, most of these studies of laryngeal function in healthy and pathological voices have achieved a certain degree of success, particularly in that pathological speakers can be broadly distinguished from healthy speakers by perturbation analysis. Further research is needed in this area, however, to refine the techniques in order to improve the

descriptive power of waveform perturbation parameters. The next chapter addresses this issue.

CHAPTER 5

THE PERTURBATION MEASUREMENT SYSTEM

CHAPTER 5

THE PERTURBATION MEASUREMENT SYSTEM

5.0 INTRODUCTION

In Chapter 3, a detailed discussion was presented for a parallel processing pitch detection algorithm. The parallel processor is an automatic system which operates on the time domain speech waveform to produce measures of fundamental frequency and amplitude. It was also demonstrated in Chapter 3 that the parallel processor is a suitable algorithm for pitch detection in a number of speech processing applications. In this chapter, the discussion will concentrate on one application of the data produced by the parallel processing PDA, as input to a system for evaluating perturbations found within F0 and A0 contours. The perturbation measurement system, to be discussed in detail in this chapter, consists of three notable features including:

1) Input data consisting of F0 and A0 contours extracted from samples of connected speech by the automatic parallel processing PDA,

2) Perturbation measurement based on excursions (i.e. deviations) of the individual input values from a local smoothed trend line of samples — a non-linear smoother consisting of a 5-point running-median plus 3-point Hanning filter produces the smoothed trend line and

3) Output parameters comprised of long-term distributional measures of frequency and amplitude perturbation based on the excursions from the smoothed trend line.

The measurement system also produces long-term distributional measures of the fundamental frequency based on the smoothed trend line of F0 values. This chapter is a detailed description of the algorithms used for evaluating the long-term parameters of frequency and amplitude perturbations as well as intonational parameters.

SECTION 5.1 -- ALGORITHMS FOR MEASURING WAVEFORM PERTURBATIONS IN F0 AND A0 CONTOURS

In Chapter 3 above, details were presented for an algorithm used to detect pitch periods from a time domain representation of the speech signal. The application of the parallel processing PDA to a sample of connected speech produces two outputs including 1) an F0 contour consisting of the inverse values of the detected pitch periods in units of Hz and 2) an A0 contour whose sample values are based on peak amplitudes derived from each detected pitch period represented in the F0 contour. Each of these contours is characterized by slow-moving intonational changes in value as well as rapid perturbatory movements which are short-term in nature. Both types of contour behavior are to be examined in order to determine their usefulness in differentiating pathological speakers from the population of healthy speakers. As can be seen from the literature review of perturbation studies in Chapter 4 above, very few studies have examined waveform perturbations displayed in samples of connected speech. The following sections describe a set

of algorithms which examine the F0 and A0 contours for long-term measures of fundamental frequency and perturbation behavior using trend line analysis techniques.

A given sample of connected speech demonstrates a number of characteristics which must be considered when designing a system which evaluates the presence of waveform perturbations within the data. Firstly, F0 and A0 contours derived from connected speech demonstrate relatively long-term changes in value which are related to the intonational aspects of the utterance. These long-term movements of F0 or A0 can be seen as rising values, falling values, regions of constant values, and change-overs between any of these states (e.g. a rising then falling F0 contour). Statistical evaluation of the long-term intonational movements of an F0 contour will demonstrate an average and range of values which characterize that particular utterance. Therefore, two speakers may be differentiated by their long-term intonational behavior, for example, where one speaker demonstrates a higher average and wider range of F0 as compared to another speaker. Secondly, at a segmental level, F0 and A0 contours derived from connected speech evidence regions of voicing (as designated by detected pitch in this system) and regions of no voicing. It is desirable when inspecting voiced segments within a contour that the onsets and offsets of voicing be reasonably well-preserved since these moments are related to gross changes in phonatory efficiency of the larynx. The last two characteristics of note are evidenced within the voiced segments in an F0 or A0 contour. There are also slow-moving but relatively short-term changes in a contour related to phonatory behavior such as vibrato. In the system to be described below, these slow-moving

changes are excluded from the measurement of the very short-term rapid movements of frequency and amplitude perturbation.

The system for evaluating F0 and perturbation parameters is designed with these various F0 and A0 contour characteristics in mind. The following discussion begins with the notion of a smoothed trend line extracted from the original input F0 or A0 contour -- this trend line retains the intonational and segmental characteristics of the input contour. A number of investigations of waveform perturbation have used a trend line approach and in this study, a simple non-linear smoothing algorithm developed by Rabiner et al. (1975) is described as a method for deriving smoothed trend lines of F0 and A0 values. For perturbation analysis, the trend line provides a useful baseline from which deviations of the unsmoothed input values from their equivalent smoothed values may be measured. The system for measuring these deviations is designed to limit the effects of the short-term slow-moving changes of F0 and A0 within voiced segments as well as normalize for the differing input levels of F0 and A0 found within a given speaker's utterance and between different speakers' voice samples. A number of long-term measures of intonation and perturbation are then presented which use the trend line approach. An additional perturbation measure is described based on Hecker and Kreuls' (1971) Directional Perturbation Factor and derived directly from the unsmoothed F0 and A0 contours.

SECTION 5.2 — NON-LINEAR SMOOTHING TO PRODUCE F0 AND A0 TREND LINES

As noted in the literature review above (Sections 4.1.3, 4.1.4 and 4.2.2), a number of investigations of waveform perturbations used a trend line approach to establish a smoothed contour from which frequency and amplitude perturbations can be measured. The trend lines were usually produced by a linear smoother, most often in the form of a running-average technique. Linear smoothing of pitch contours produced useful results since most of these studies evaluated sustained vowel phonations in which most of the variation in period duration and amplitude was due to perturbations and slower-moving changes such as vibrato. However, linear smoothing may not be completely appropriate for creating trend lines of frequency and amplitude contours due its low-pass filtering characteristics. As a low-pass filter, the linear smoother will fail to bring errant data points back into a trend line (Rabiner and Schafer 1978) — these errant points may be either actual large perturbations of the waveform or gross errors in pitch extraction (e.g. the tracking of the second harmonic within the input speech signal). In the running-average approach a number of local measures of F0 or A0 will be distorted with the insertion of a single errant sample into the linear smoother. In addition, linear smoothing will not be appropriate for the development of trend lines from contours of pitch and amplitude extracted from stretches of connected speech. The low-pass filtering effects of the linear smoother will severely distort the sharp discontinuities in the contours which represent the transitions from voiced to unvoiced segments and vice versa. In the spectral sense, these transitions contain high-frequency energy which will be smeared by linear smoothing (Rabiner et al. 1975). The following sections discuss the basic properties of the non-linear smoother as set out by Rabiner et al. (1975) as well as

specific implementation issues.

Linear and non-linear smoothers differ in their basic approaches to signal filtering. The approach of linear smoothing is the separation of non-overlapping frequency components within a signal. In non-linear smoothing, the basic approach is to separate components within a signal which are characterized as being either smooth or rough (i.e. noise-like). For example, perturbation analysis of F0 contours consists of separating the noise-like movements of F0 (i.e. the perturbations) from the overall smooth trend line movement of the contour which is related to intonation in the case of connected speech. A given signal x(n) is treated as the combination of its smooth and rough components such that:

$$x(n) = S[x(n)] + R[x(n)]$$

where S[x(n)] is the smoothed part of the signal and R[x(n)] represents the rough component. No ideal non-linear smoothing mechanism exists which can completely separate the smooth and rough portions of a given signal. A reasonable first approximation to the desired properties of non-linear smoothing is produced by the use of a running-median. The running-median of a signal ($MED_L[x(n)]$) is defined simply as the median value of L samples $x(n),\ldots,x(n-L+1)$. For a low order running-median containing an odd number of samples, the samples are ordered in value and the center value is chosen as the median for that set of samples.

The running-median displays a number of useful properties for the smoothing of various types of speech signals. Firstly, a running-median will not smear out sharp discontinuities in the signal, as long as the duration of a discontinuity exceeds some critical duration. Therefore, it is possible to preserve realistic discontinuities in F0 contours, in particular, transitions from voiced to unvoiced states and vice versa. Figure 5.1 displays some examples of smoothing using linear and non-linear techniques. Figure 5.1a is the input signal and it has been smoothed by a linear smoother (5.1b -- 5-point running-average), and 2 different running-medians (5.1c -- 3-point median; 5.1d -- 5-point median). Note that in the case of the 2 signals produced by the medians that the sharp discontinuities in the signals have been preserved. However, the sharp discontinuities of the input signal have been smeared by the linear smoother as seen in Fig. 5.1b. The choice of the length of the running-median is strictly dependent on the minimum duration of the discontinuity which is to be preserved in the signal. In the present study, voiced (i.e. F0 values greater than 0 Hz) and unvoiced (i.e. F0 values equal to 0 Hz) segments of an F0 contour are operationally defined as those segments consisting of three or more sequential F0 values of either state. Therefore, a median filter with a duration of 5 samples is required to preserve discontinuities of 3 samples or more.

The second useful property of the running-median is closely related to the first property discussed above. That is, the running-median inherently smoothes out sharp discontinuities in the signal which are shorter than the critical duration of the filter. Very short discontinuities in the F0 contour are considered to be

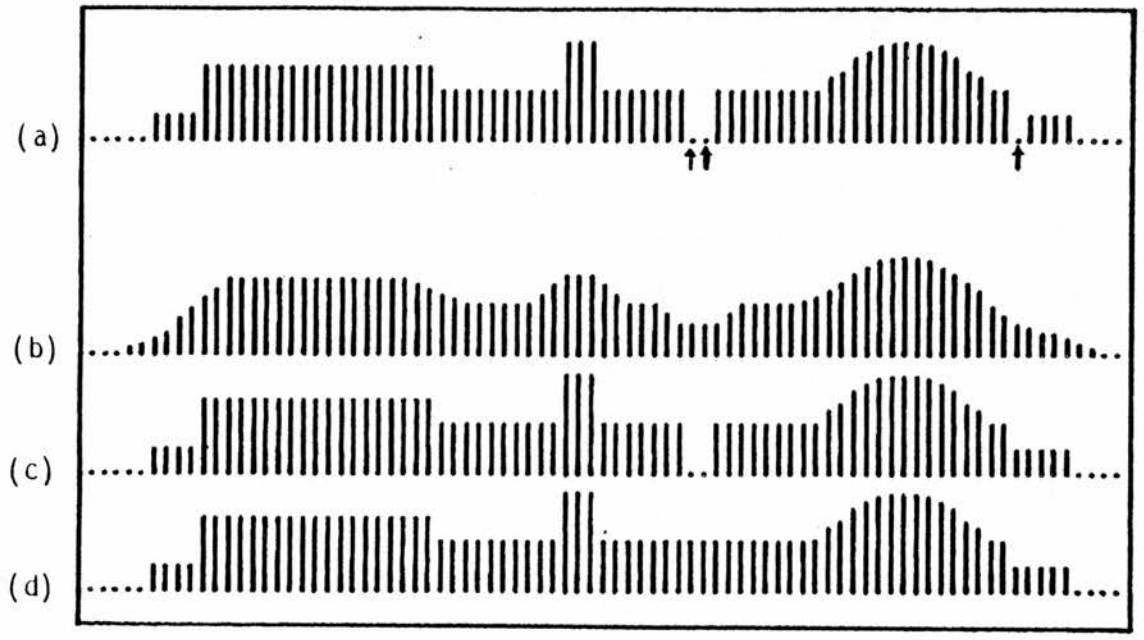


Figure 5.1 Some test examples of smoothing using linear and non-linear techniques. (a) input contour displaying sharp transitions, zero value data points (located at the arrows) and a smooth rise/fall; (b) after linear smoothing with a 5-point running-average smoother; (c) after non-linear smoothing with a 3-point running-median smoother; (d) after non-linear smoothing with a 5-point running-median smoother. (After Rabiner et al. 1975).

gross errors produced by the pitch detection algorithm or actual large perturbations of the signal produced during phonation. Linear smoothing would distort the overall shape of a smoothed F0 contour by averaging these gross F0 measures with the other local F0 values. Figure 5.1 demonstrates the effects of non-linear and linear smoothing of short sharp discontinuities in the input contour. The short discontinuities of particular interest are the sequences of 1 and 2 zero values in the input waveform highlighted by the arrows. The 3-point running-median of Fig. 5.1c smoothes out the discontinuity of length 1 while the 5-point running-median of Fig. 5.1d also smoothes out the discontinuity consisting of 2 samples. Note in Fig. 5.1b that a general local distortion of the contour can be seen at the location of the 2 sample discontinuity which has been caused by the linear smoothing. As a result of the operational definition for the critical duration of voiced and unvoiced segments within F0 contours, a 5-point median will smooth out all short discontinuities of 1 and 2 samples. The evaluation of these short discontinuities for the purpose of perturbation analysis will be discussed in detail in the section on perturbation parameters below.

The last useful property of the running-median to be discussed here is its ability to approximately follow low-order polynomial trends evidenced in speech signals. F0 and A0 contours show polynomial-like trends which are related to the intonational movements of a signal such as a change from rising to falling pitch. The ideal trend lines for perturbational analysis should have all the intonational movements preserved within it from which perturbatory movements can be evaluated. Figure 5.1a displays a rise/fall within the input contour which is in fact a quadratic

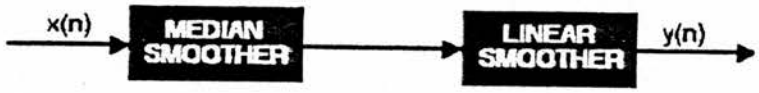


Figure 5.2 Block diagram of simple non-linear smoothing system. The additional linear element consists of a 3-point Hanning window. (After Rabiner et al. 1975).

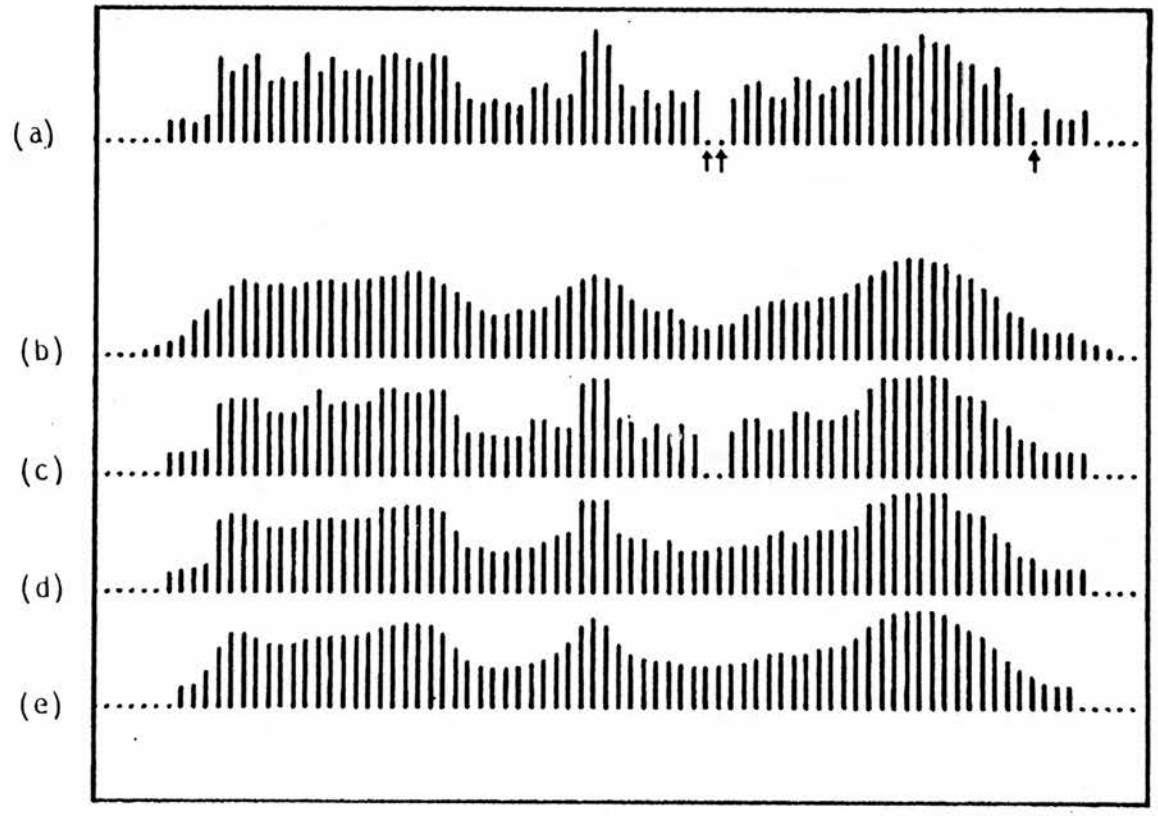


Figure 5.3 Further test examples of smoothing using linear and non-linear techniques. (a) same input contour as in Fig. 5.1a with the addition of broadband noise; (b) after linear smoothing with a 5-point running-average smoother; (c) after non-linear smoothing with a 3-point running-median smoother; (d) after non-linear smoothing with a 5-point running-median smoother; (e) after the combined smoothing of a 5-point running-median smoother and 3-point Hanning window. (After Rabiner et al. 1975).

polynomial of length equal to 15 samples. Note that all 3 smoother outputs in Fig. 5.1b-d demonstrate reasonable approximations to the original quadratic polynomial of the input signal though the running-medians appear to have produced less smearing of the signal as compared to the output of the linear smoother. Thus, the 5-point running-median chosen for this study will produce reasonable output for intonational and segmental aspects of F0 and A0 contours.

The running-median is very good at preserving sharp above-threshold discontinuities and eliminating very short ones in a given speech signal. However, median smoothing alone does not provide sufficient smoothing of the smaller noise-like components often seen in speech signals. For perturbation analysis, these small movements of a contour are the waveform perturbations which the filter system was originally designed to remove from the F0 and A0 data. A reasonable compromise in smoothing effects can be reached by including an element of linear smoothing along with the non-linear smoothing of the running-median. Since the median filter does provide some smoothing of the small noise components in a signal, a fairly low-order linear smoother can be used to complete the smoothing process. Rabiner et al. recommend a 3-point Hanning window for the linear smoothing which is a symmetrical finite impulse response filter with the following impulse response characteristics:

$$\begin{aligned} h(n) &= .25 \quad n=0 \\ &= .50 \quad n=1 \\ &= .25 \quad n=2 \end{aligned}$$

where the h(n) are the coefficients used for the linear filtering. Figure 5.2 displays a block diagram of the simple smoothing algorithm consisting of a running-median and Hanning window filters. The output of this algorithm y(n) is an approximation to the

smoothed component S[x(n)] of the input signal. Figure 5.3 displays the results of adding broadband noise to the input signal of Fig. 5.1a. In this example, median smoothing alone has performed an inadequate job of filtering out the broadband noise as can be seen in the rough outputs of Figs. 5.3c and d (the 3-point and 5-point running-medians respectively). The output of the linear smoother as seen in Fig. 5.3b demonstrates the usefulness of linear filters for removing the small broadband noise components from the signal. The linear smoother outputs of Figs. 5.1b and 5.3b are very similar in appearance. A good compromise in smoothing is provided by the combination of median and Hanning filters as seen in Fig. 5.3e. In this output, the discontinuities displayed in the input are fairly well preserved and the broadband noise has been smoothed to a reasonable degree. One practical issue in implementing the non-linear smoothing algorithm is generating the initial and final values of the output which are outside the range of the filters. Briefly, these points are set to zero values and therefore no attempt is made to extrapolate these points. The implications of this issue for perturbation analysis will be discussed in detail in the following sections.

It should be noted that various implementations of the non-linear smoother described above have been used in studies investigating other areas of speech processing. In an investigation of a speaker-independent digit-recognition system, Sambur and Rabiner (1975) applied a non-linear smoother to a number of parameters prior to machine recognition of spoken digits. The parameters included average zero-crossings count, the normalized error from LPC analysis, the total energy and the pole frequencies

extracted by LPC analysis. Non-linear smoothing was used for these parameters since the rough portions of the original contours reflect the use of a set analysis interval length for speech parameters which were dynamic in character. Sambur (1978) applied a non-linear smoother to pitch contours used in a system for canceling noise in speech signals.

Figure 5.4 displays the effects of applying the non-linear smoother to an F0 contour output by the parallel processing PDA for a small section of connected speech uttered by a healthy speaker. In each part of Fig. 5.4, fundamental frequency is plotted in units of Hz along the ~~abscissa~~ ordinate while the ~~ordinate~~ abscissa displays the order of the F0 values. Each F0 contour in Fig. 5.4 consists of an order of 200 samples which is approximately 1.7 sec. of speech data. The input F0 contour (5.4a) displays a number of characteristics. Firstly, there is an overall downward movement of fundamental frequency associated with the long-term intonational aspects of the utterance. Secondly, the F0 contour is segmented into regions of voicing (i.e. values greater than 0 Hz) and regions of no voicing (i.e. values equal to 0 Hz). Within the voiced portions of the contour, there are small movements of F0 which are the waveform perturbations of interest. Finally, there are sections of short sharp discontinuities at the offsets of the voiced sections which may be the result of the PDA attempting to detect pitch data from low amplitude voiced data. These short discontinuities are considered to be gross errors of pitch extraction due to the irregularity of the speech signal. Figures 5.4b and c are the F0 contours produced by the application of simple running-medians of 3 and 5 points to the input F0 contour. Note that in each contour,

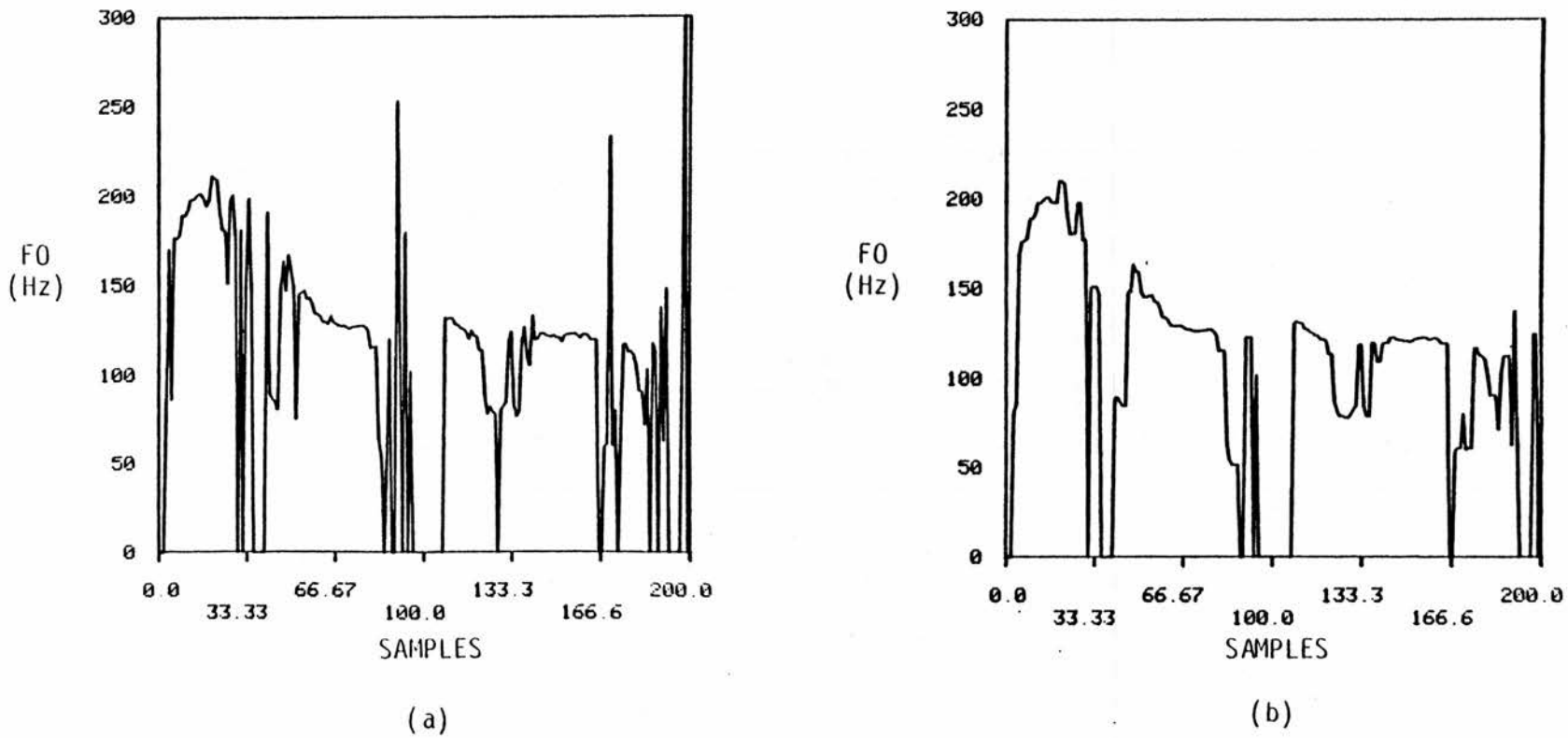


Figure 5.4 Some examples of applying the non-linear smoother to an F0 contour extracted from a sample of connected speech produced by a healthy male speaker (the contour is an order of 200 values, approximately 1.7 sec). (a) F0 contour extracted by the PDA; (b) after non-linear smoothing with a 3-point running-median smoother; continued on next page.

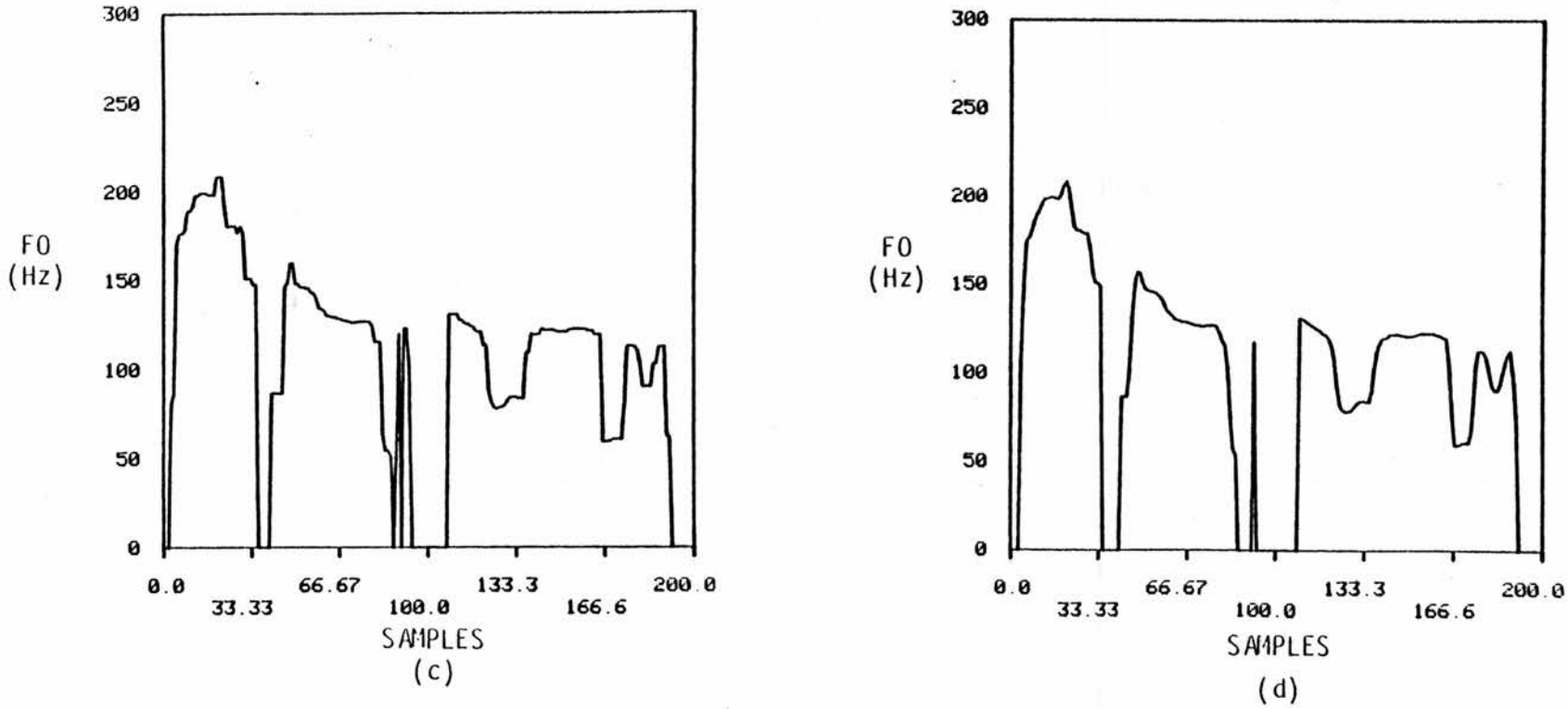


Figure 5.4 Continued. (c) after non-linear smoothing with a 5-point running-median smoother; (d) after the combined smoothing of a 5-point running-median and 3-point Hanning window.

median smoothing has preserved sharp discontinuities associated with the onsets and offsets of voicing while also eliminating very short discontinuities in regions of rapid and gross fluctuations of the F0 values (at voicing offsets). The effect of the longer median of 5 points can be seen by comparing Fig. 5.4b to 5.4c where the 5-point median has brought more of the F0 values into the overall contour. However, neither running-median completely smoothed out the finer perturbatory movements evidenced in the input F0 contour. Figure 5.4d displays the results of applying the simple smoothing algorithm (see Fig. 5.2) to the input F0 contour. The addition of the Hanning window as a linear smoother has produced a much smoother contour from which perturbation measures can be made. As can be seen in Fig. 5.4d, a useful compromise in F0 contour smoothing is provided by a combination of linear and non-linear smoothing.

Most of the discussion thus far has been concerned with the smoothing of F0 contours for trend line analysis of pitch perturbations. The simple smoothing system is also applied to the A0 contour for amplitude perturbation analysis which uses a trend line approach as well. Figure 5.5 displays the smoothing of an A0 contour which has been derived from the peak amplitudes of the pitch periods used to produce the F0 contour of Fig. 5.4a. In each part of Fig. 5.5, amplitude is displayed on the abscissa while the order of the A0 values is plotted along the ordinate (the duration of samples is approximately 1.7 sec as in Fig. 5.4). Firstly, it is interesting to note that most of the erratic behavior of the input F0 contour of Fig. 5.4a is located in regions of very low amplitude as seen in Fig. 5.5a. The effects of median smoothing of the input A0 contour can be seen in the output contours of Figs. 5.5b and c.

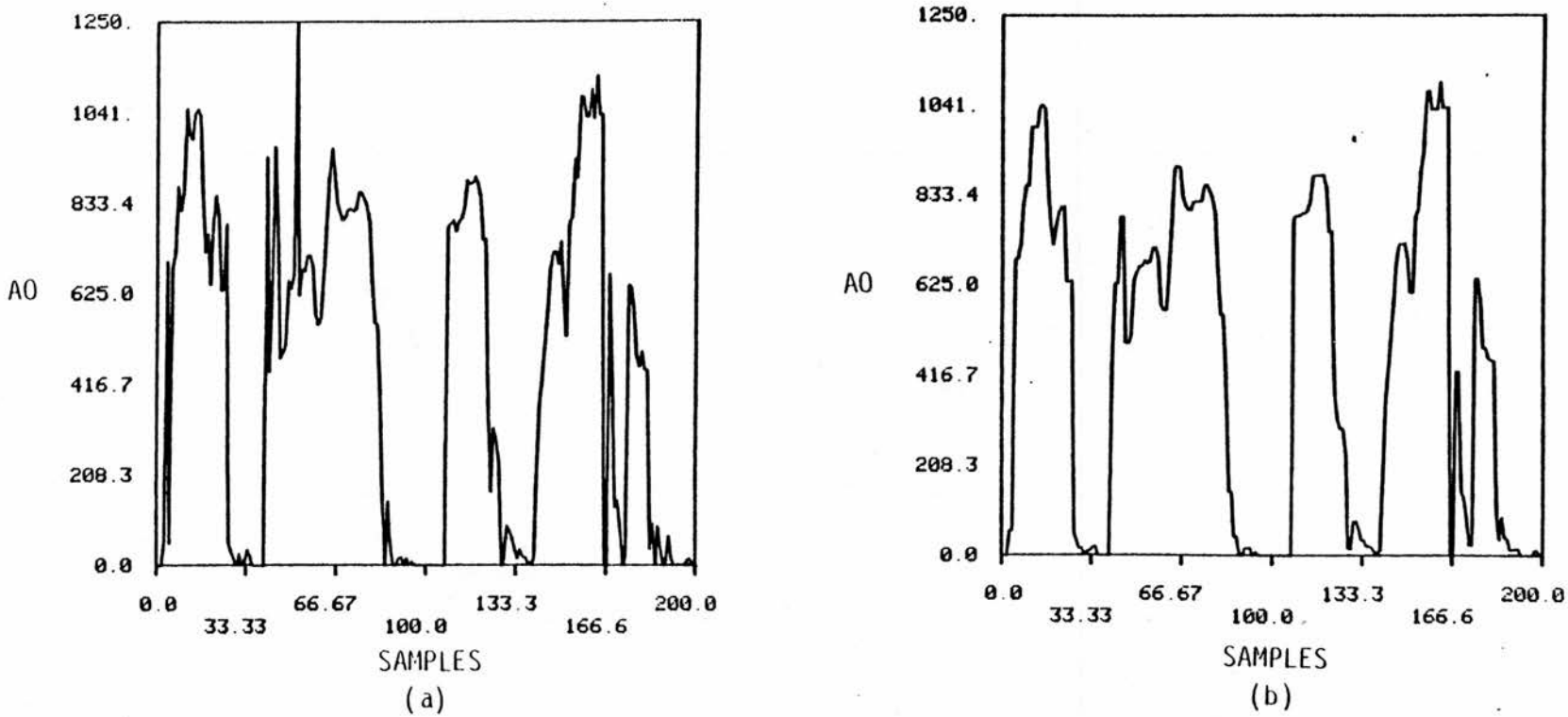
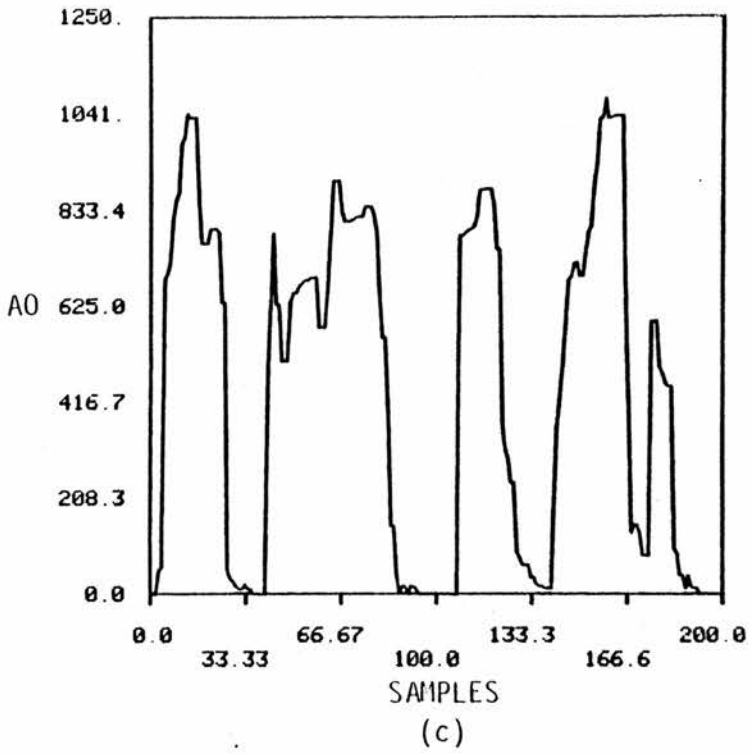


Figure 5.5 Some examples of applying the non-linear smoother to an AO contour extracted from a sample of connected speech produced by a healthy male speaker (the contour is an order of 200 values, approximately 1.7 sec). (a) AO contour extracted by the PDA; (b) after non-linear smoothing with a 3-point running-median smoother; continued on next page.



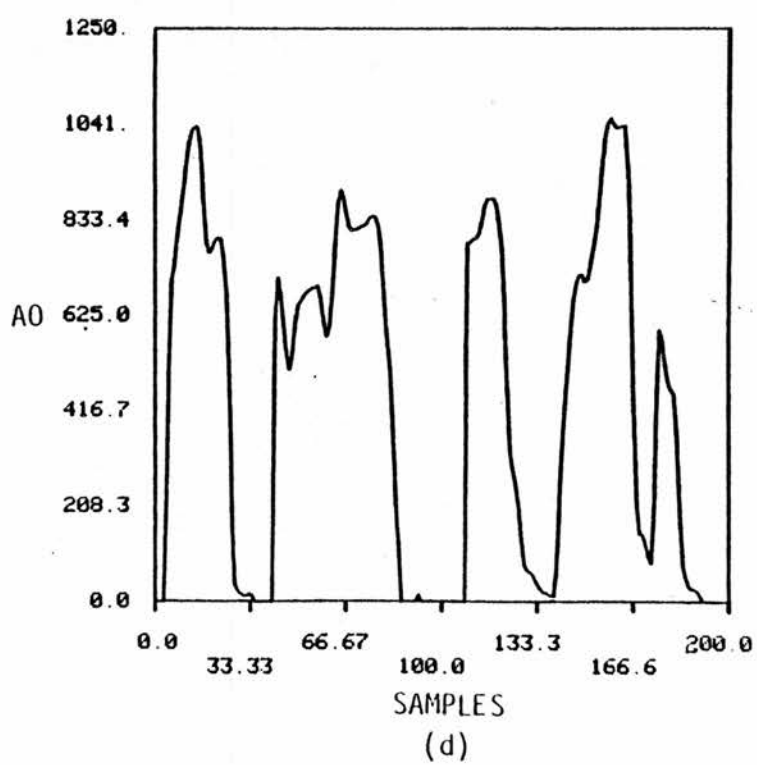


Figure 5.5 Continued. (c) after non-linear smoothing with a 5-point running-median smoother; (d) after the combined smoothing of a 5-point running-median and 3-point Hanning window.

As was noted for Fig. 5.4, the median smoothing does a very good job of preserving sharp discontinuities associated with the segmental aspects of the contour while also eliminating short discontinuities. The smoothed A0 contour of Fig. 5.5d is a good example of the usefulness of the combined smoothing of a running-median and Hanning window for the amplitude data.

Figures 5.6 and 5.7 display the effects of non-linear smoothing for F0 and A0 contours derived from an utterance produced by a speaker with carcinoma of the vocal folds. Figure 5.6 is the F0 contour while Fig. 5.7 is the A0 contour (the ordinate displays 233 A0 values which have been derived from approximately 1.7 sec. of speech). The input F0 contour in Fig. 5.6a displays a considerable amount of erratic behavior particularly for the first half of the contour which is associated with perturbed phonation and gross pitch detection errors. As in the case of the healthy speaker, much of the erratic behavior occurs in regions of low amplitude as seen in the A0 contour of Fig. 5.7a. The usefulness of the combined median and Hanning smoothers for smoothing of the waveforms and preserving sharp discontinuities can be seen in Figs. 5.6d and 5.7d for the F0 and A0 contours, respectively.

SECTION 5.3 -- EXCURSIONS FROM THE TREND LINE

The application of the non-linear smoother to a given F0 or A0 contour produces a trend line from which perturbations can be measured for connected speech samples. The basic unit of perturbation measurement is defined as the difference between an input unsmoothed value of F0 or A0 and its equivalent smoothed value along the trend line derived by the non-linear smoother. This basic

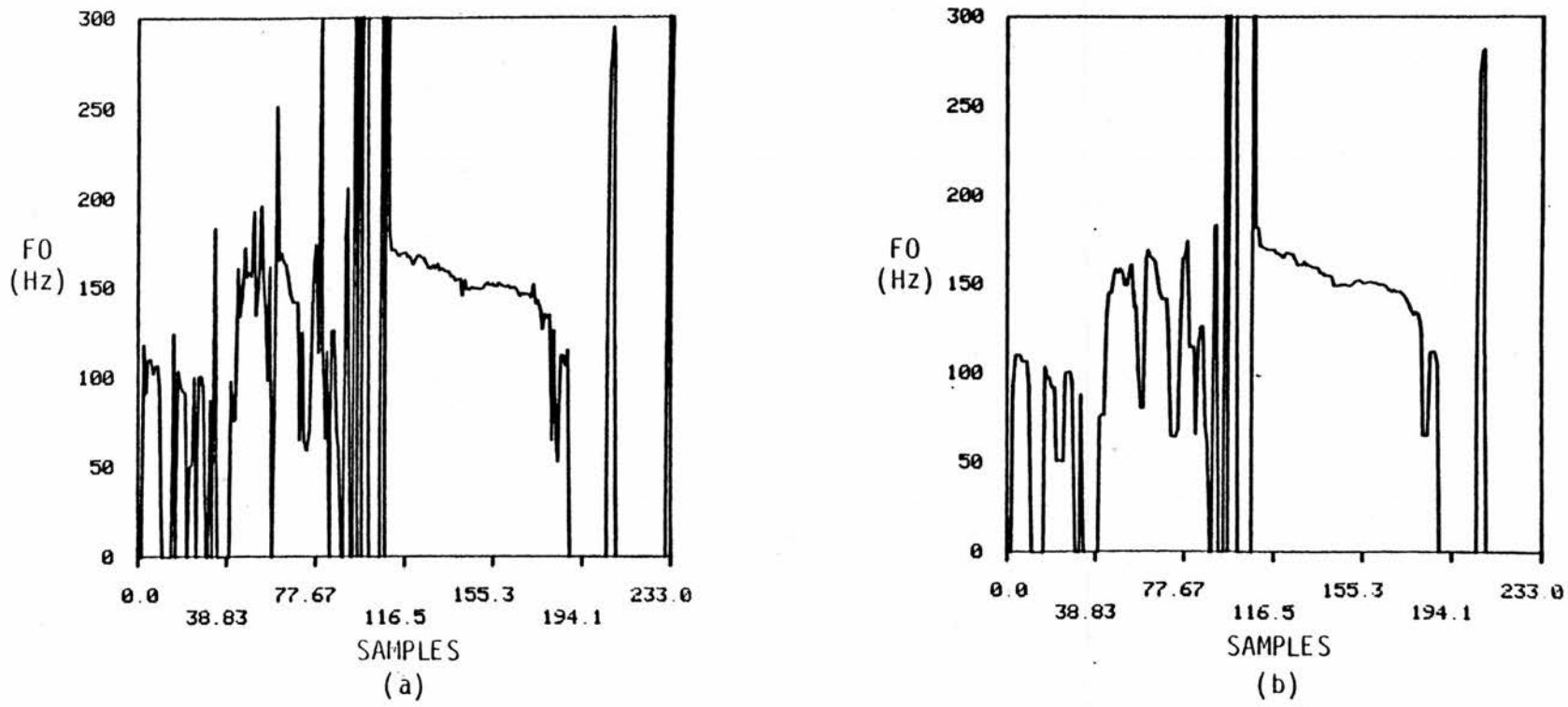
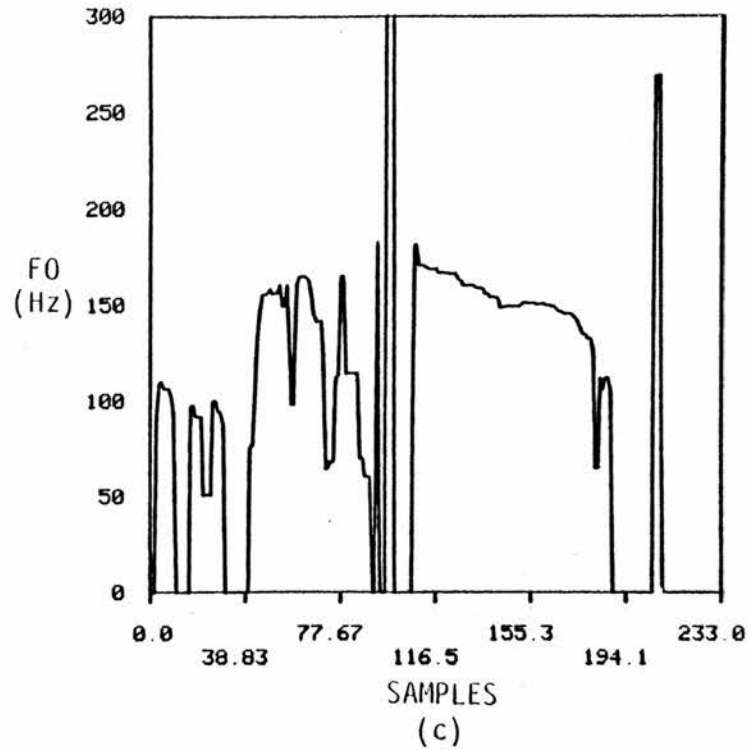


Figure 5.6 Some examples of applying the non-linear smoother to an F0 contour extracted from a sample of connected speech produced by a pathological male speaker (the contour is an order of 233 values, approximately 1.7 sec). (a) F0 contour extracted by the PDA; (b) after non-linear smoothing with a 3-point running-median smoother; continued on next page.



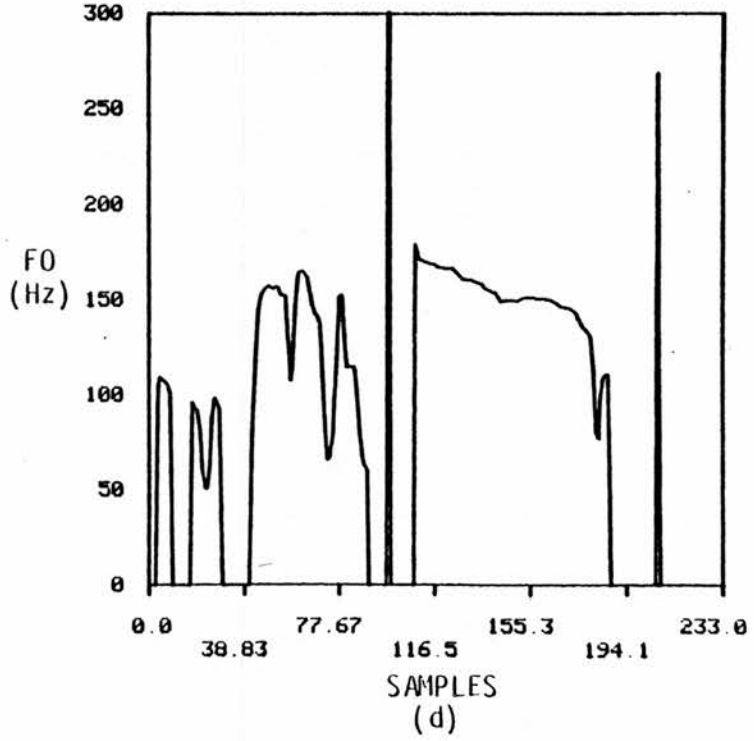


Figure 5.6 Continued. (c) after non-linear smoothing with a 5-point running-median smoother; (d) after the combined smoothing of a 5-point running-median and 3-point Hanning window.

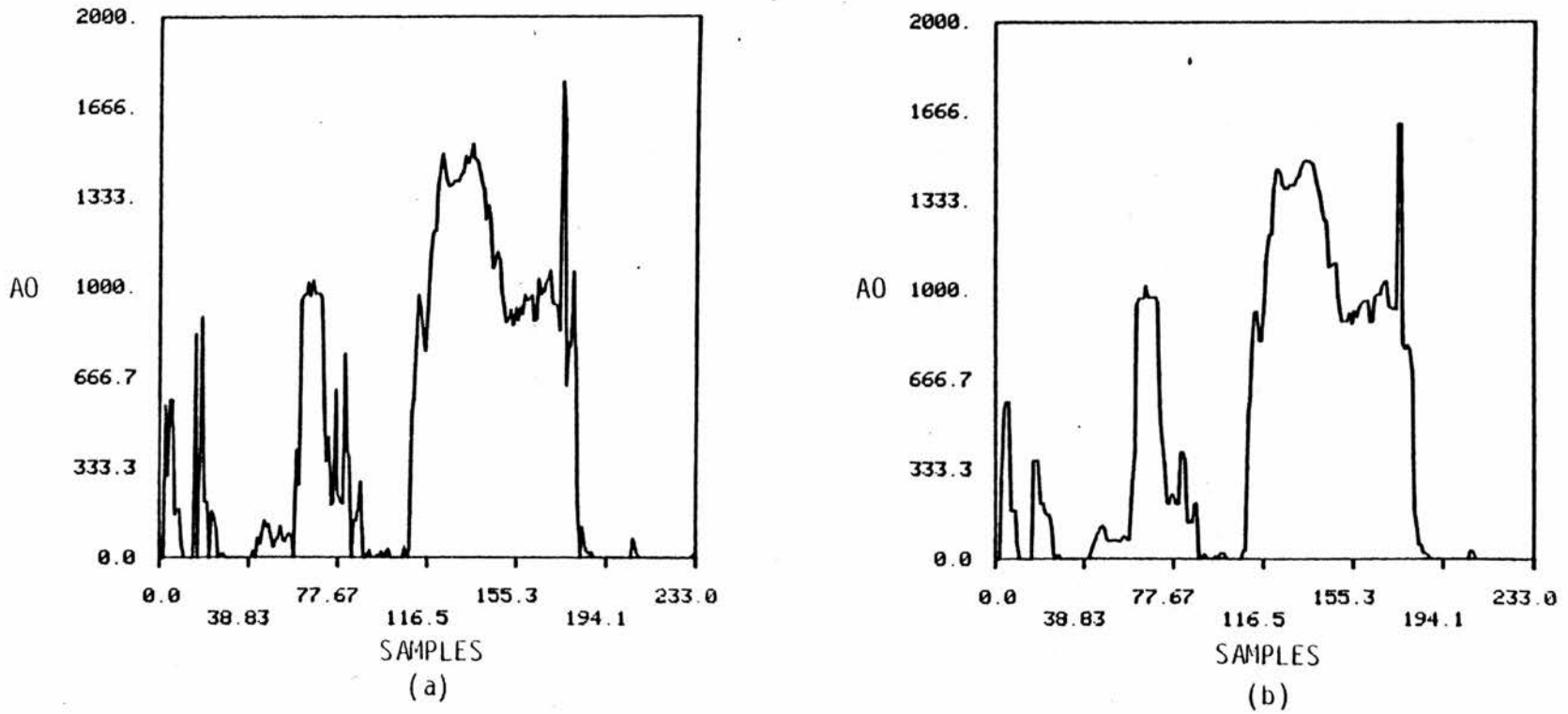


Figure 5.7 Some examples of applying the non-linear smoother to an AO contour extracted from a sample of connected speech produced by a pathological male speaker (the contour is an order of 233 values, approximately 1.7 sec). (a) AO contour extracted by the PDA; (b) after non-linear smoothing with a 3-point running-median smoother; continued on next page.

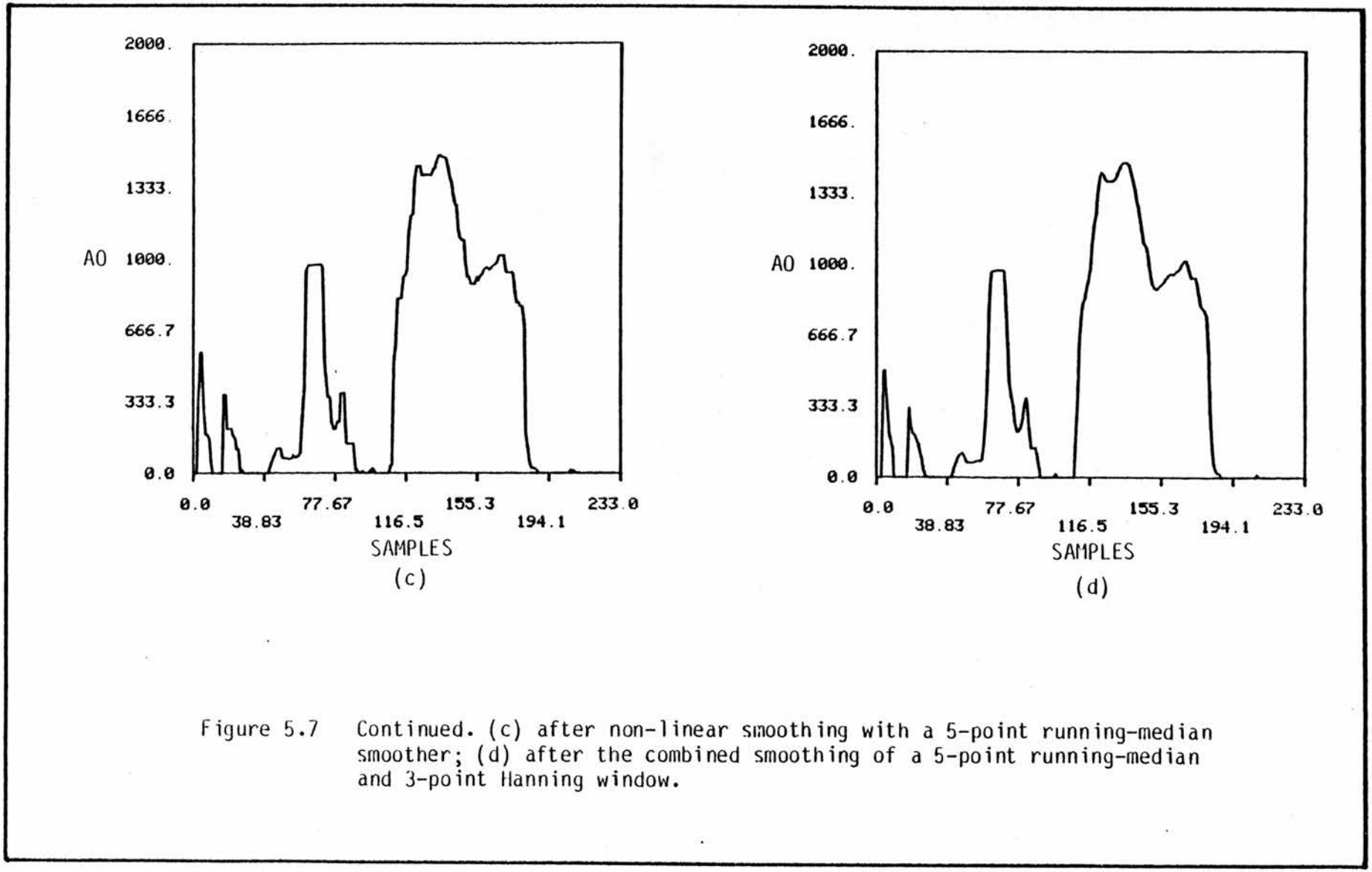


Figure 5.7 Continued. (c) after non-linear smoothing with a 5-point running-median smoother; (d) after the combined smoothing of a 5-point running-median and 3-point Hanning window.

unit of perturbation is called an Excursion -- this unit is measured in relation to a smoothed trend line in order that slow-moving modulations and intonational movements of F0 and A0 are excluded from contributing to the estimation of perturbation parameters. Figure 5.8 displays a block diagram for implementing the system for measuring excursions using the non-linear smoother of Fig. 5.2. As can be seen from this diagram, the signal representing the excursions e(n) is derived by subtracting the original input values of F0 or A0 from their equivalent smoothed samples y(n). The signal e(n) is equivalent to the rough portion R[x(n)] of the input signal. The correct implementation of the system shown in Fig. 5.8 requires that the proper delay of the input signal be used prior to computing the difference between the smooth and input signal values. For the present system, the running-median of 5 points has a delay of 2 samples while the 3-point Hanning window has a delay of 1 sample. The total delay for this simple non-linear smoother is 3 samples.

An excursion is derived for each output of the non-linear smoother and defined as the difference between the unsmoothed input value of F0 or A0 and its equivalent smoothed sample. Each excursion of F0 is stored in 4 formats: 1) signed excursion in Hz — the difference between input and smoothed F0 in units of Hz with the algebraic sign retained, 2) signed excursion in percent (SE%) -- the ratio of the signed excursion in Hz to its equivalent smoothed F0 value multiplied by 100 (SE% = e(n)/y(n)*100), 3) magnitude of excursion in Hz -- the absolute value of the signed excursion in Hz and 4) magnitude of excursion in percent (ME%) -- the absolute value of the signed excursion in percent. These four formats are also produced for a given input A0 contour with the appropriate amplitude

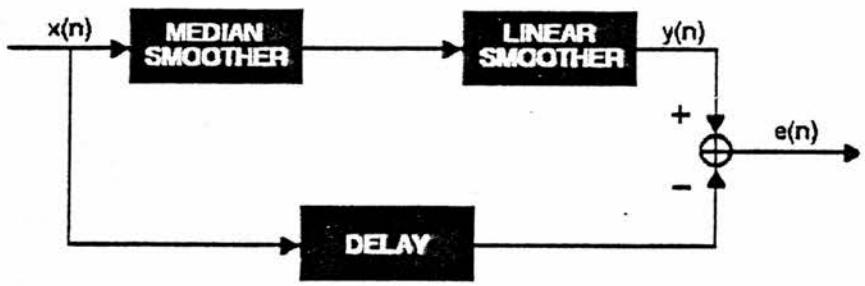


Figure 5.8 Block diagram of smoothing system used for measuring excursions e(n). The median smoother is 5 samples in duration and the linear smoother is a 3-point Hanning window; the resultant delay is equal to 3 points.

values. Of the 4 formats, the SE% and ME% should be the most revealing for perturbation analysis since they are normalized versions of excursion. These measures should be normalized with respect to varying levels of F0 and A0 evaluated 1) within a sample of continuous speech produced by a single speaker and 2) between different speakers. The normalization is necessary since a number of studies (see, for example, Lieberman 1961; 1963; Smith and Lieberman 1969; Hollien et al. 1973; Koike 1973; Koike et al. 1977; Smith et al. 1978; Horii 1979; 1980; Murray and Doherty 1980; Sorenson et al. 1980; Wilcox and Horii 1980; Benjamin 1981; Ramig and Ringel 1983; Ludlow et al. 1983a) noted a positive relationship between unnormalized values of perturbation and average pitch period of a voice sample. That is, as average pitch period duration increased for a given voice sample (i.e. as F0 decreased), the average perturbation value increased. Therefore, speakers who use a low range of fundamental frequencies produce increased perturbation values compared to speakers with higher average fundamental frequencies.

At this point, it is appropriate to discuss those instances in which excursions are not derived for a given F0 or A0 contour. Firstly, no excursions are estimated for any unvoiced segment which is defined as a sequence of 3 or more zero values in a given F0 contour. Secondly, there are instances when samples within a smoothed trend line are undefined due to delays associated with the operation of the running-median (delay = 2 samples) and the Hanning window (delay = 1 sample) components of the non-linear smoother. The total system delay of 3 samples for the non-linear smoother means that the initial and final 3 endpoints of any given smoothed

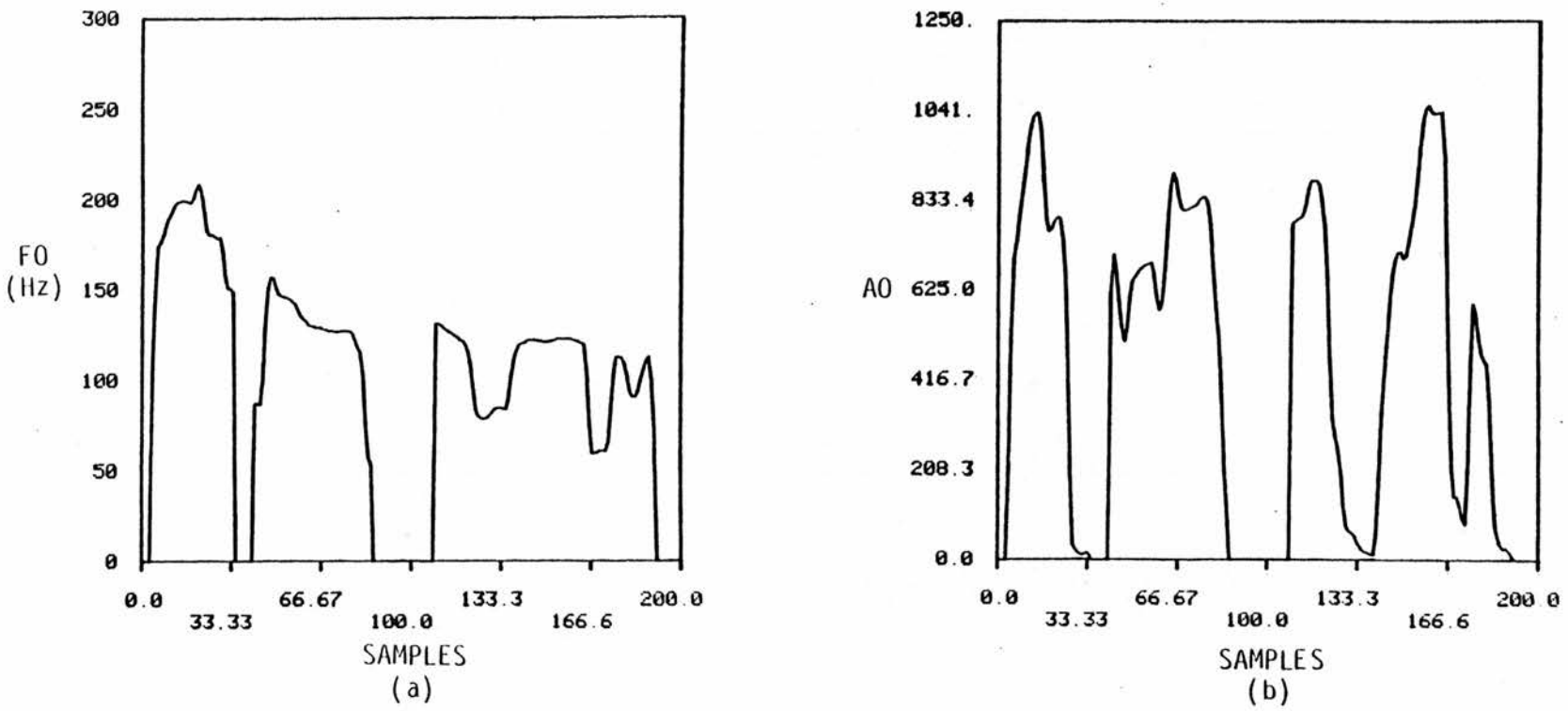


Figure 5.9 Examples of F0 and A0 trend lines produced by the non-linear smoother when boundaries are set for acceptable values of F0 (limits: 40 to 240 Hz for males; 75 to 450 Hz for females). (a) F0 trend line -- equivalent to trend line presented in Fig. 5.4d for the healthy male speaker; (b) A0 trend line -- equivalent to trend line presented in Fig. 5.5d for the healthy male speaker.

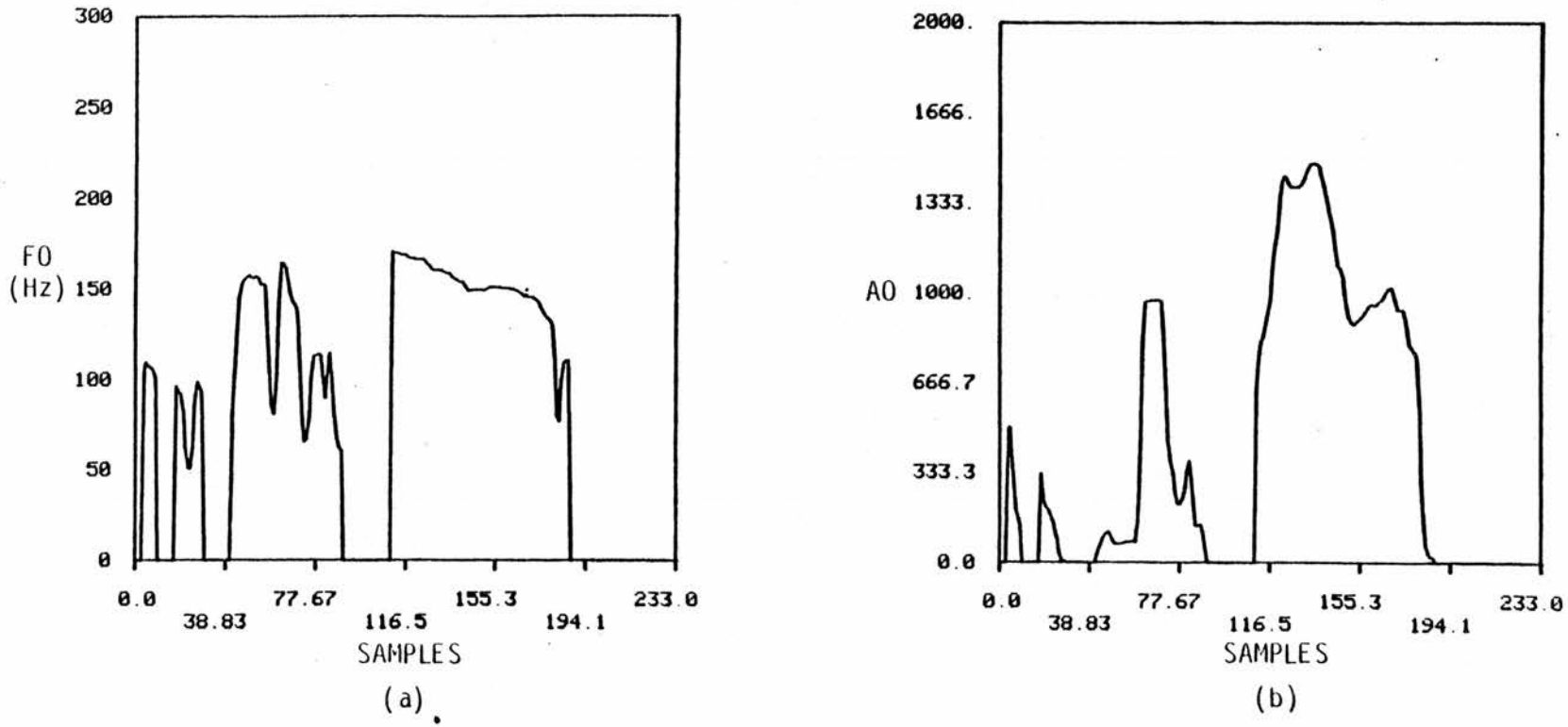


Figure 5.10 Examples of F0 and A0 trend lines produced by the non-linear smoother when boundaries are set for acceptable values of F0 (limits: 40 to 240 Hz for males; 75 to 450 Hz for females). (a) F0 trend line -- equivalent to trend line presented in Fig. 5.6d for the pathological male speaker; (b) A0 trend line -- equivalent to trend line presented in Fig. 5.7d for the pathological male speaker.

contour are undefined. In these instances, no attempt is made to extrapolate these missing values in the smoothed contour; rather these values are set to zero and no excursion measures are completed. The third instance in which no excursion measure is completed is when zero values appear in the Hanning window for linear smoothing. Zeros within the Hanning window are the result of unvoiced sections of a given contour which has been preserved by the running-median component of the smoothing algorithm. Therefore, smearing of the data at voicing onsets and offsets will be produced by the linear smoothing effects of the Hanning window. That is, unvoiced F0 values equal to zero are averaged with non-zero values of F0 to produce a smeared contour sample. In this case, the conservative approach has been taken where no excursions are measured whenever a zero value appears in the Hanning window — it is not certain what the true excursion will be for an input F0 or A0 value which has been smeared by the linear smoothing. In addition, the actual output of the smoothing algorithm is set to zero such that smearing of the filtering sort does not appear in a contour. Fourthly, excursions are not determined for short, sharp discontinuities of 1 or 2 samples within voiced segments of a given F0 contour. In particular, short discontinuities with values of F0 equal to 0 Hz would produce excursions of 100% (i.e. SE% = [(x-0)/x]*100 = 100% where x is any smoothed contour value) leading to uncertainty in the resultant perturbation results. However, it is recognized that short discontinuities are related to irregularities in the original input speech signal which have resulted in momentary errors in pitch extraction. These short discontinuities of 1 or 2 samples within voiced segments are called Anomalies and the number of anomalies in a given F0 contour is

totaled as a special measure of perturbation.

One further issue related to the measurement of excursions is discussed here. It was decided to limit gross errors of pitch period detection from the measurement of excursions by the use of limits of acceptable F0 values. These gross errors are eliminated by setting bounds on the acceptable range of fundamental frequencies which can be input to the smoothing algorithm. These bounds of acceptability are sex-specific such that acceptable F0 values derived by the parallel processing PDA ranged from 40 to 240 Hz for male speakers and 75 to 450 Hz for female speakers. F0 values outside these specific frequency ranges are set to zero prior to being input to the smoothing system. Therefore, the measurement of excursions will be effected in two ways by this boundary system. Firstly, if a sequence of F0 values is longer than the critical duration for the non-linear smoother (i.e. 3 samples or more) and all the values are outside the pre-set limits for acceptable measures, then that sequence is equivalent to an unvoiced segment of the contour and no excursion measures are completed. Secondly, if the F0 sequence is shorter than the critical duration and outside the range of acceptability for F0, then the sequence is equivalent to a short discontinuity -- in this case, its occurrence is noted in the total count of anomalies found for a given contour and no excursion is measured. The effects of setting the limits of acceptable F0 values for excursion analysis can be seen in Figs. 5.9 and 5.10. In Fig. 5.9 (the healthy speaker), the smoothed trends of F0 (5.9a) and A0 (5.9b) have been produced with the use of boundaries set from 40 to 240 Hz. Figure 5.9a is equivalent to the F0 contour presented in Fig. 5.4d for the healthy speaker while

Fig. 5.9b is equivalent to the A0 contour presented Fig. 5.5d. Note that the high F0 values located at approximately 100 samples has been removed from Fig. 5.9a due to the setting of the limits. This sample is treated as unvoiced since it appears in a segment of erratic values (see Fig. 5.4a) which are effectively treated as zeros. It should be noted that bounds for acceptable F0 values also control the smoothing of A0 contours. That is, if an F0 value is found to be out of limits then the associated A0 value is also considered out of bounds. This effect can be seen in Fig. 5.9b where the value at approximately location 100 has also been removed due to the unacceptability of its associated F0 sample. The use of limits for F0 data can be clearly seen in Fig. 5.10 of the pathological speaker. The F0 contour of Fig. 5.10a is equivalent to the contour in Fig. 5.6d while the A0 contour of Fig. 5.10b is related to Fig. 5.7d. A certain number of F0 values have been removed by the F0 boundaries, in particular, as can be seen in the comparison between the F0 contours in Figs. 5.10a and 5.6d.

In summary, there are a number of instances in which excursion measures are not completed for a given F0 or A0 contour. Of particular interest are those instances in which input F0 and A0 values are present but discarded which results in a loss of information regarding perturbatory behavior. A conservative approach has been taken here and therefore a small number of samples are lost at onsets and offsets of voicing (due to the Hanning window) and in regions of gross F0 values which are outside the limits of acceptable measures. It is assumed that long-term measurement of F0 and A0 contours will produce sufficient data to enable full evaluation of waveform perturbations. Further research

should be completed to determine the acoustic phenomena associated with the onsets and offsets of voicing — these phonatory actions may provide a good deal of information about the presence of disorders in the vocal folds. In the present measurement system, a number of samples from onsets of voicing are lost by the PDA as well as the excursion measurement scheme.

A more detailed description of the nature of trend line excursions is given in this section. Figure 5.11 displays F0 contour data which has been taken from Figs. 5.6a and 5.10a of the pathological speaker. In Figure 5.11a, the input F0 contour (designated by plus signs) and its equivalent smoothed trend line (marked by the solid line) have been plotted for samples 40 to 100 (this is approximately .44 sec of speech data) to enlarge the contours for a closer inspection of excursions. A number of details can be seen in Fig. 5.11a for the two related contours. Firstly, there is the overall intonational movement of the data highlighted by the smoothed F0 trend line. Secondly, this is a very perturbed section of the F0 contour as revealed by a number of large excursions of the input F0 values from the smoothed trend line. Thirdly, a number of the input samples appear as short discontinuities of the input signal (these samples are encircled to highlight them). These short discontinuities are either values of zero or outside the upper limit for acceptable F0 values (i.e. 240 Hz). The short discontinuities within the range of approximately 40 to 85 samples which are brought into the trend line by the running-median, are not evaluated for excursions but here added to the total count of anomalies for this F0 contour. The final sequence of F0 values from sample 89 onwards is all treated as unvoiced (i.e. set to zero

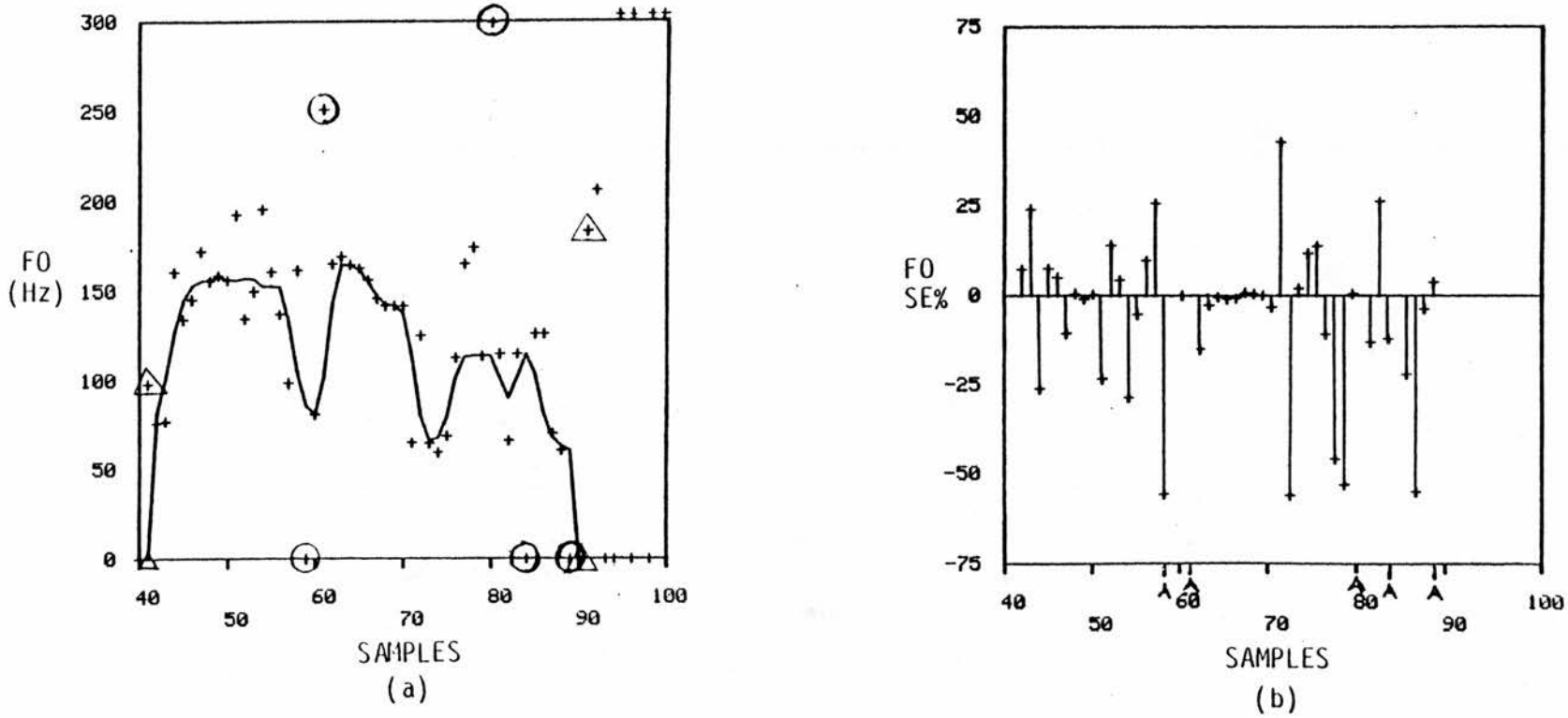
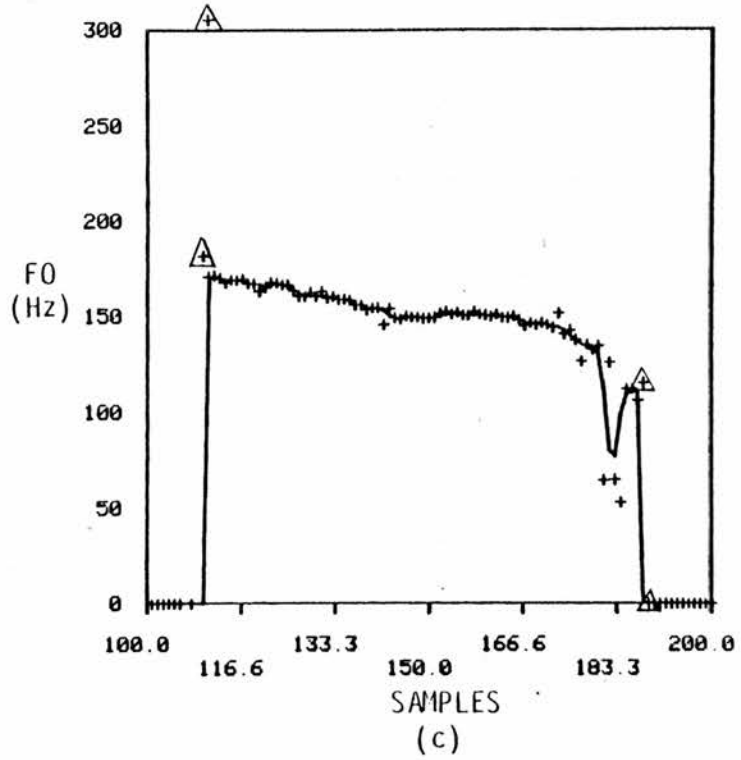


Figure 5.11 Two examples of measuring signed excursions in percent from F0 contours extracted from samples of connected speech produced by a male pathological speaker. Example 1 -- a highly perturbed F0 contour: (a) comparison of F0 contour (+) and equivalent smooth trend line (solid line); (b) F0 signed excursions in percent derived from the comparisons in (a); Circles -- anomalous F0 values in contour; triangles -- endpoints of voiced segment not measured for excursion values; A -- locations of anomalous values in parts (a) and (c); continued on next page.



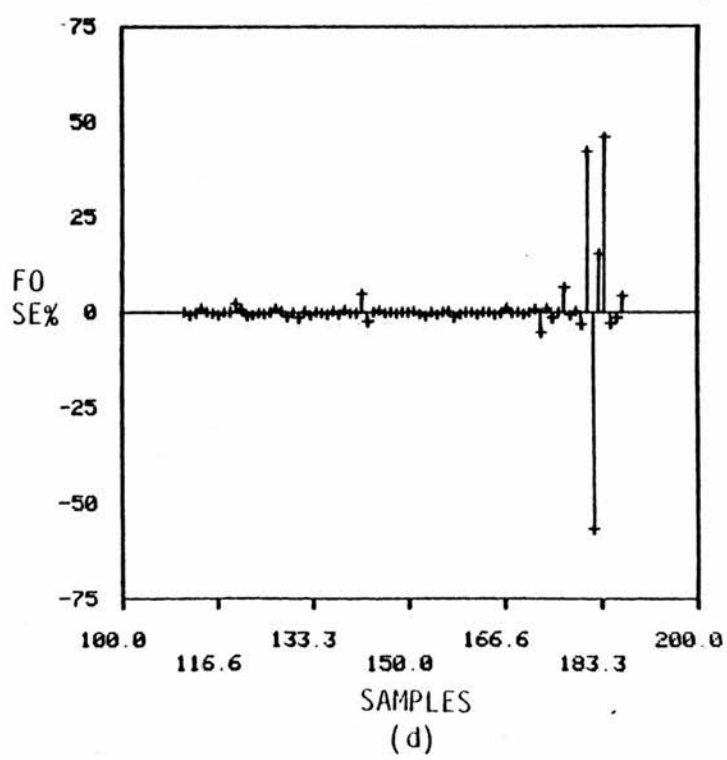


Figure 5.11 Continued. Example 2 -- a less perturbed F0 contour: (c) comparison of F0 contour and equivalent trend line; (d) F0 signed excursions in percent derived from the comparisons in (b).

values in the smoothed trend line) since it is a series of short discontinuities of zeros and out of bound values. The two within bound values beyond sample 90 are also set to zero since they are a short discontinuity of acceptable values within a sequence of zero values (this is a very special case of anomaly for which there is no measurement scheme at this time). Finally, the initial and final 2 samples of the input contour (marked by triangles) are not evaluated for excursions since their values would have been smeared by the 3-point Hanning window which also contained zero values at their onset and offset. Figure 5.11b displays the signed excursions in percent of the F0 values contained in Fig. 5.11a. The abscissa of this Figure is plotted in units of percent from -75% to 75% while the ordinate represents the order of the signed excursion values. In this Figure, all the long-term intonational movement of F0 has been removed leaving behind the normalized values of the F0 excursions (the signs of the excursions reflect the variation in the data above and below the baseline). It can be seen from this Figure that this is a very perturbed segment of the F0 contour due to a large number of substantial excursions from the baseline. These are representative of the units of excursion used for the derivation of perturbation parameters. The locations of the anomalies related to this series of excursions are marked by "A" along the ordinate of the Figure.

Figure 5.11c displays more regular segments of F0 contours derived from the pathological speaker's data shown in Figs. 5.6a and 5.10a (samples 100 to 200 which is approximately .725 sec of speech data). It can be seen in Fig. 5.11c that a majority of input F0 values are very close to their equivalent smoothed values

along the trend line which suggests a fairly regular segment of phonation. Some irregularities occur at the onsets and offsets of this section of the F0 contour. This regularity of the input F0 contour is clearly seen in Fig. 5.11d which displays the SE% derived from the 2 contours in Fig. 5.11c. A majority of SE%s are close to the baseline and a few substantial excursions occur near the end of this sequence of values.

Figure 5.12 demonstrates excursion behavior for segments of the A0 contours presented in Figs. 5.7a and 5.10b for the pathological speaker. In Fig. 5.12a, the input and smoothed A0 contours are presented for samples 40 to 100 of Figs. 5.7a and 5.10b. The overall intonational movement of the A0 data is highlighted by the smoothed trend line. It is interesting to note that many of the gross measures of F0 seen in Fig. 5.11a are located in regions of very low signal amplitude as seen in Fig. 5.12a. The deviations of the input A0 values from the smoothed trend line also suggest that this is a very perturbed segment of A0 contour though it is not quite as clear as in the equivalent F0 contour due to the magnitude of the input A0 values. The irregularity of the A0 contours becomes evident when the signed excursions in percent of the A0 values are seen in Fig. 5.12b. This Figure demonstrates many substantial excursions from the baseline, some of these excursions being greater than 100%. These A0 excursions are basic units for measuring perturbations in A0 contours. Figures 5.12c and d are the A0 contours derived from samples 100 to 200 of Figs. 5.7a and 5.10b of the pathological speaker. The input and smoothed A0 contours of 5.12c are from the region of less perturbed F0 contour movement described previously for Fig. 5.11c. The input A0 contour of Fig.

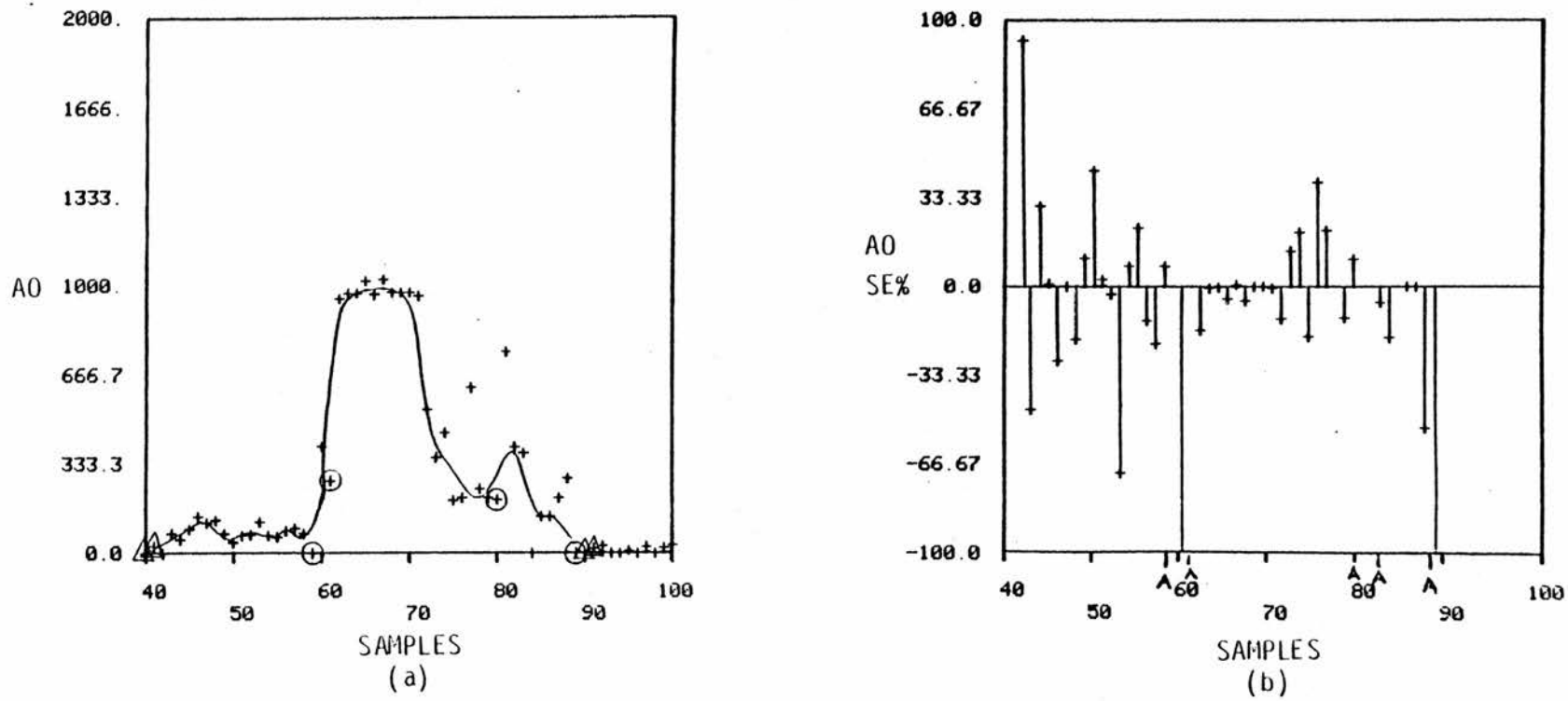
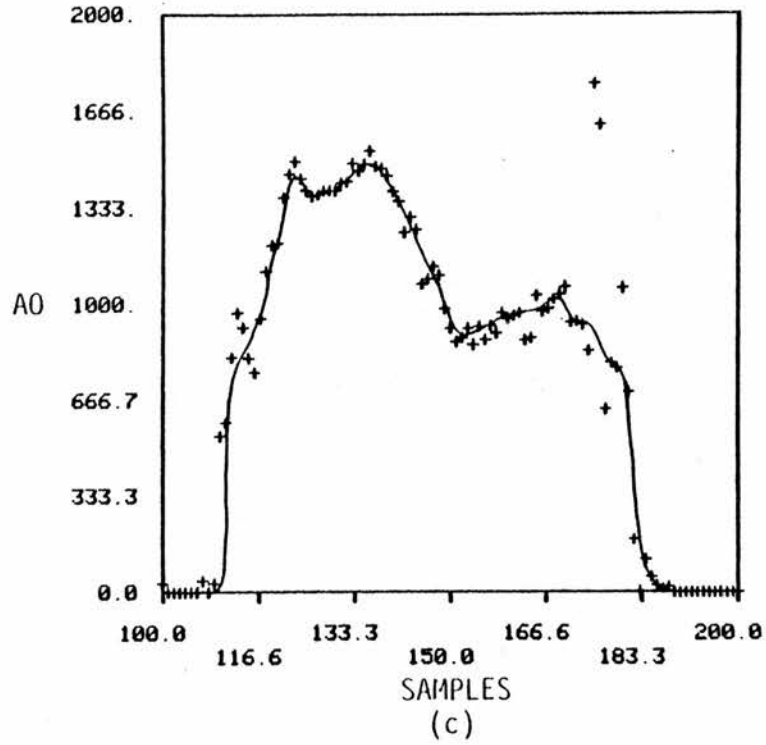


Figure 5.12 Two examples of measuring signed excursions in percent from AO contours extracted from samples of connected speech produced by a male pathological speaker. Example 1 -- a highly perturbed AO contour: (a) comparison of AO contour (+) and equivalent smooth trend line (solid line); (b) AO signed excursions in percent derived from the comparisons in (a); Circles -- anomalous AO values in contour; triangles -- endpoints of voiced segment not measured for excursion values; A -- locations of anomalous values in parts (a) and (c); continued on next page.



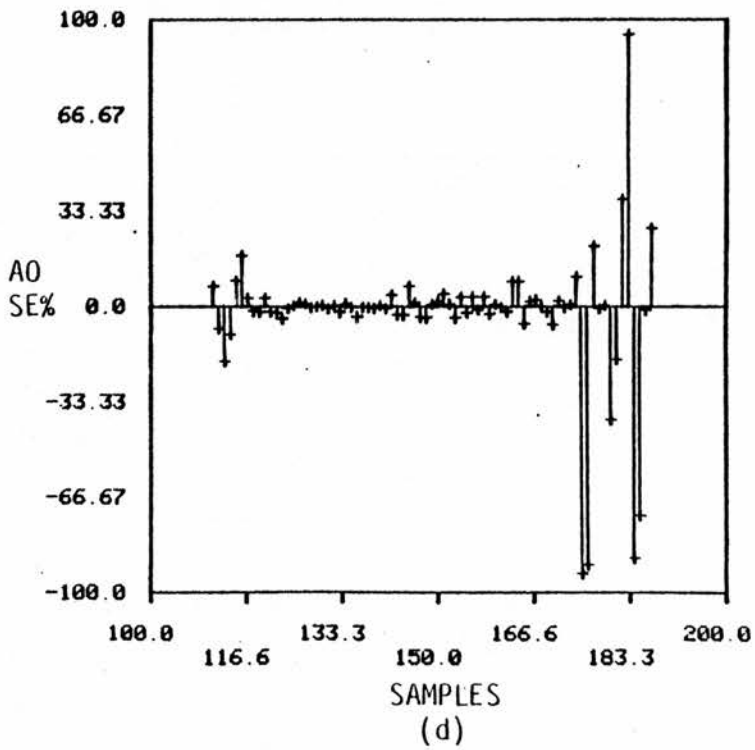


Figure 5.12 Continued. Example 2 -- a less perturbed AO contour: (c) comparison of AO contour and equivalent trend line; (d) AO signed excursions in percent derived from the comparisons in (b).

5.12c also appears less perturbed as many of the values are close to the smoothed A0 trend line. However, the input A0 contour in general does not appear to be as smooth as its equivalent F0 contour. In particular, there are a few input A0 values near the end of the contour in Fig. 5.12c which are quite far removed from the trend line. These samples may reflect instances in which the A0 values are not derived from peak maxima in the waveform (i.e. peak function M1) but rather some other peak value associated with the matching process for choosing the most likely pitch period. Thus, the pitch period synchronization problem noted for the parallel processor will affect the measurement of waveform shimmer. However, if the synchronization problem does arise due to waveform irregularity then perhaps an increase in shimmer values is one way of quantifying the irregularities. Figure 5.12d displays the signed excursions in percent derived from the two A0 contours of Fig. 5.12c. Many of the excursions are close to the baseline but a certain amount of irregularity is present. Large excursions of A0 are seen at the onsets and offsets of this contour.

SECTION 5.4 -- LONG-TERM INTONATIONAL AND PERTURBATION PARAMETERS

The simple non-linear smoother described above is incorporated into a more general program for analyzing waveform perturbations in F0 and A0 contours. Upon output, a number of parameters are stored which are useful in determining long-term intonational and perturbation parameters for a given voice sample. The outputs of the program include: 1) the smoothed trend lines for F0 and A0, 2) the excursions derived from the differences between the input and smoothed contours and 3) a number of perturbation parameters

calculated during the execution of the perturbation analysis program. The following sections describe the parameters in detail.

1. Parameters based on the Non-linear Smoother Output -- 2 sets of long-term parameters are based on the output of the simple non-linear smoother including a) intonational parameters and b) perturbation parameters.

a. Intonational Parameters -- Several measures of the overall behavior of the smoothed F0 contour as derived by the non-linear smoother are evaluated for each voice sample. It is expected that these long-term F0 measures will reflect the typical level and range of F0 produced for each voice sample -- these measures will reflect certain aspects of the intonation used to produce each voice sample as uttered by healthy and pathological speakers. In addition, certain disorders of the vocal folds affect the mass and stiffness of the tissues in the folds and it is expected that these pathological voices may be detected by long-term intonational parameters which are substantially greater or lesser in value as compared to speakers with healthy larynges (Mackenzie, Laver and Hiller 1983). Two intonational parameters of particular interest are derived by distributional analysis of the smoothed F0 values output from the non-linear smoother including:

F0-AV -- The average fundamental frequency in units of Hz for all the non-zero F0 samples in the smoothed contour; the use of the smoothed contour means that values outside pre-set limits (40-240 Hz males, 75-450 Hz females) have been eliminated from the original data pool to limit the effects of gross measures in F0 produced by

the PDA.

F0-DEV -- The standard deviation of the fundamental frequency in units of Hz for all the non-zero values in the smoothed F0 contour. This parameter represents the typical range of F0 produced by each speaker for a given speaking task. The pre-set limits for acceptable values are also used for producing this parameter.

Other long-term distributional parameters are derived for the intonational aspects of the smoothed F0 contour including the total range (i.e. the minimum and maximum F0 values), the median and modal fundamental frequency in units of Hz. At this time, equivalent long-term intonational parameters are not derived for the smoothed A0 contour.

Two histograms of smoothed F0 values are displayed in Figure 5.13. Figure 5.13a contains F0 data from the speech of the healthy male speaker presented in the previous figures in this section. Figure 5.13b shows the F0 values derived from the voice sample of the pathological speaker also used thus far in this section. In both cases, 40 seconds of speech were analyzed in order to produce the histograms. The values in the bins of each histogram range from 40 to 240 Hz which are the limits of acceptable F0 values set for male speakers. The distribution of the data for the healthy speaker in Fig. 5.13a appears to be normally distributed (though it should be noted that the normality of the distribution has not been statistically tested). An F0-AV of 112.4 Hz and F0-DEV of 21.9 Hz was found for this distribution. The histogram of Fig. 5.13b for the pathological speaker appears to be slightly skewed to the lower

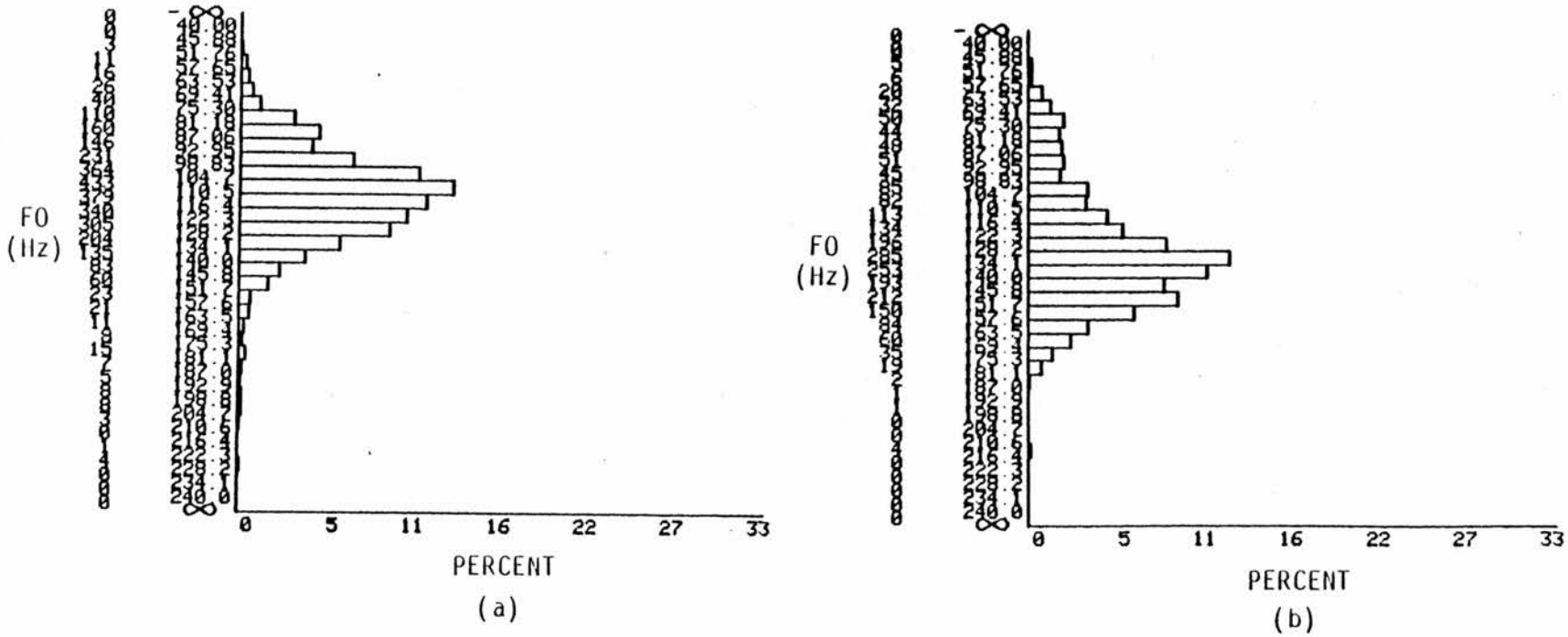


Figure 5.13 Histograms of F0 values present in F0 trend lines derived from 40 secs of connected speech produced by (a) the healthy male speaker and (b) the pathological male speaker. The intonational parameters F0-AV and F0-DEV are taken from this type of histogram.

values of F0 as compared to Fig. 5.13a though the overall differences between the two histograms do not appear to be great. The pathological speaker evidenced an F0-AV of 128.3 Hz and F0-DEV of 25.8 Hz for this histogram. Though both F0-AV and F0-DEV are greater for the pathological speaker as compared to the healthy speaker, it remains to be seen from further evaluations in the following chapter if the differences are significant.

b. Perturbation Parameters -- A number of perturbation parameters are based on the excursions derived from the smoothed trend line produced by the smoothing system. For any perturbation measure, it is expected that certain levels of waveform perturbation will be found which are typical of voice samples produced by speakers with healthy voices. For many pathological conditions of the voice, in particular where an asymmetrical disruption of the vocal fold tissues is present, it is expected that the resultant long-term excursion-based perturbation measures will be substantially different compared with the healthy speakers (Mackenzie et al. 1983). It is felt that the perturbation measures with the greatest potential for detecting and differentiating pathological speakers are based on the excursions formats which include normalized percentage factors, that is, the signed excursion in percent and the magnitude excursion in percent. Their greater potential arises from the normalization factor within their computation which limits differences in overall F0 and A0 level within the voice sample of a speaker as well as between samples produced by different speakers. A number of perturbation parameters are derived by distributional analysis of the ME% and SE% values produced by the excursion analysis including:

J-AVEX/S-AVEX -- These parameters represent the average excursion (AVEX) as found for the magnitude excursions in percent for F0 (jitter) and A0 (shimmer) contours, respectively. The magnitude of the excursion is used as the input to this parameter in order that all non-zero values of F0 or A0 will produce a non-zero mean value (i.e. excursions vary around a zero baseline and therefore the absolute value of the excursions seems more appropriate here.) If there is a tendency towards the production of larger than average waveform perturbations in a given voice sample then these 2 parameters will reflect this behavior.

J-DEVEX/S-DEVEX -- These parameters represent the standard deviation of the excursions (DEVEX) as derived for the SE% for F0 (jitter) and A0 (shimmer) contours, respectively. In this case, signed excursions are the input since it is expected that a normal-like distribution of excursions will occur around a mean value close to zero. If there is a tendency towards the production of larger than average excursions of F0 and A0 then these should be revealed as a larger spread of excursion values as seen in the DEVEX measures for jitter and shimmer.

Two histograms of the signed excursions in percent for F0 data are displayed in Figure 5.14. Figure 5.14a contains the SE% data derived from the speech sample of the healthy speaker while Fig. 5.14b presents the SE% data for the pathological speaker. In both cases, 40 seconds of speech was used to produce these histograms. The values in the bins of each histogram range from -100% to 100%. Two points should be noted for these bin values. Firstly, the absolute range of the data has been limited to +/- 100% only for the

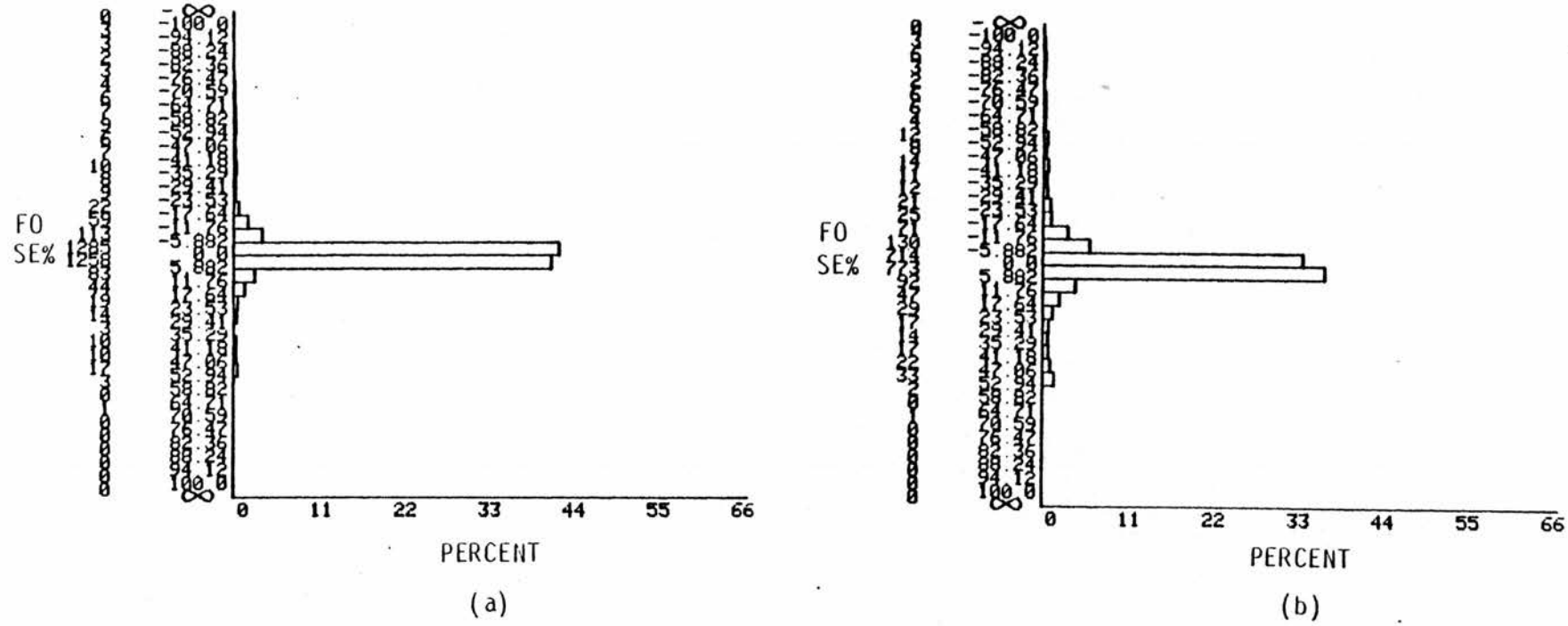


Figure 5.14 Histograms of frequency signed excursions in percent values derived from 40 secs of connected speech produced by (a) the healthy male speaker and (b) the pathological male speaker. The frequency perturbation parameter J-DEVEX is taken from this type of histogram.

purposes of this presentation in order that the distribution of a majority of SE% values in a given sample are clearly visible. In reality, the limits are much wider to permit the inclusion of large excursions from the smoothed trend line which are greater than 100%. Secondly, the values contained within the bins reflect the nature of the excursions such that a) negative values exist due to the excursions below the smoothed trend line and b) percentage values are used to limit contour movements within and between given voice samples. The histograms of SE% values are used to derive the perturbation parameter J-DEVEX in this example — it is expected that the distribution of the signed excursions around a zero mean will produce a better representation of the standard deviation. It can be seen that the distribution of SE%s for the healthy speaker in Fig. 5.14a appears to be more peaked and narrow compared to the pathological speaker's SE% distribution in Fig. 5.14b. That is, the healthy speaker produced smaller excursions of jitter and more often than the pathological speaker. The J-DEVEX for the healthy speaker is 16.6% and the pathological speaker evidenced a J-DEVEX of 20.6%. It remains to be seen if a significant difference exists between these two speakers for the J-DEVEX value.

Figure 5.15 displays the two histograms of SE% data derived for the A0 values found for the voice samples of the healthy and pathological speakers. The values of the bins of the histograms range from -100% to 100% -- as explained above, these limits are solely for the purpose of enlarging the distribution for convenient observation. Figure 5.15a is the distribution of SE% values of A0 for the healthy speaker and Fig. 5.15b is the distribution derived from the pathological speaker's voice sample. The differences in

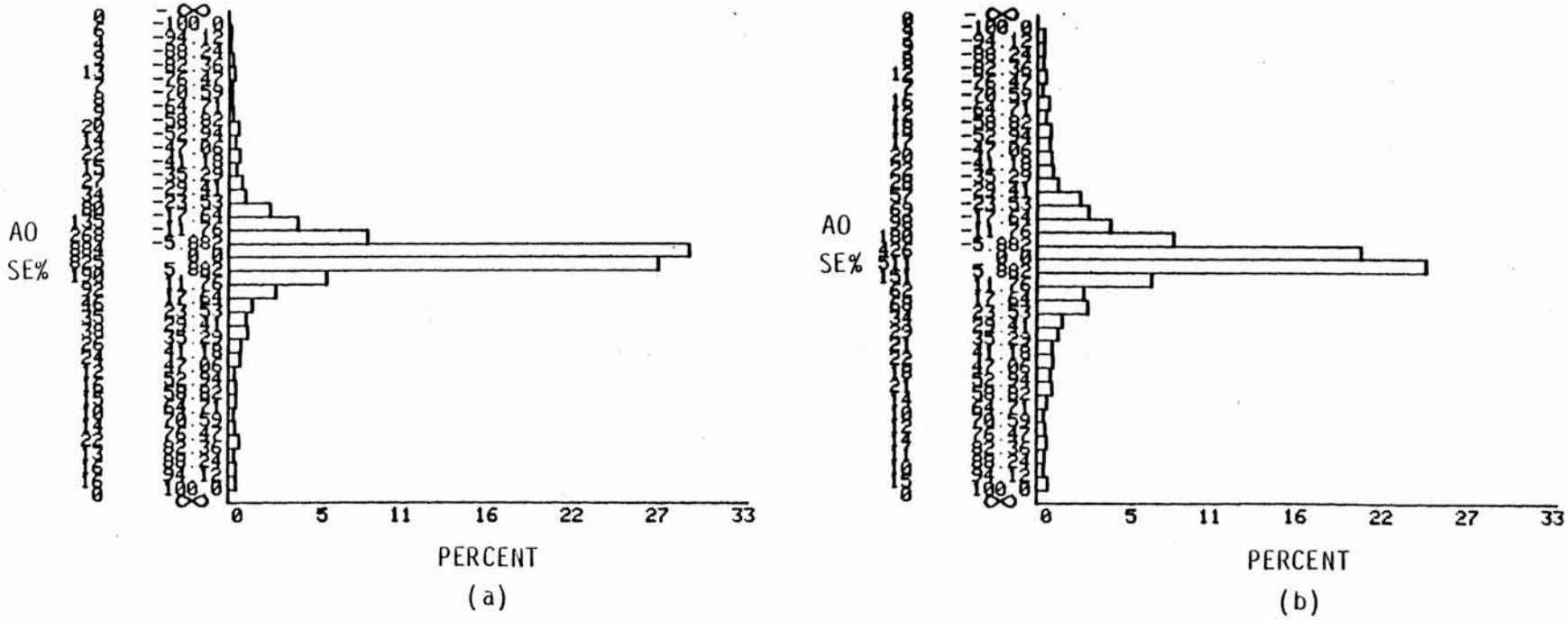


Figure 5.15 Histograms of amplitude signed excursions in percent values derived from 40 secs of connected speech produced by (a) the healthy male speaker and (b) the pathological male speaker. The amplitude perturbation parameter S-DEVEX is taken from this type of histogram.

distribution between these two histograms are not quite as clear as found in Fig. 5.14 though the healthy speaker's distribution appears to be slightly more peaked and narrower as compared to the histogram of the pathological speaker. That is, the healthy speaker produced smaller excursions of A0 and more often than the pathological speaker. The perturbation measures S-DEVEX is derived from this type of histogram -- the healthy speaker evidenced an S-DEVEX of 58.9% while the pathological speaker evidenced an S-DEVEX of 77.8%.

Other distributional measures such as range, median and mode are also calculated and stored for the SE% and ME% parameters.

J-RATEX/S-RATEX -- These parameters represent the rate of excursions (RATEX) found in an F0 and A0 contour respectively. RATEX is the percentage of points in a given contour where a magnitude of excursion in percent is greater than a pre-set threshold. The pre-set threshold is used to quantify the number of significant perturbations in any given voice sample (similar to Lieberman's 1963 minimum threshold used to produce the Perturbation Factor). The pre-set threshold is set to 3% in the present study, because even in the healthiest voices, uttering a sustained monotone vowel, the successive pitch periods typically show approximately 2% frequency jitter, in a normal distribution (Hanson 1978). The 3% threshold enables one to discount this factor from the J-RATEX measure. The choice of threshold for S-RATEX is less certain since equivalent statistical statements are not available for shimmer. A 3% threshold was also chosen for determining S-RATEX for the present study. The following RATEX values were found for the voice sample

of the healthy speaker: J-RATEX = 23.73% and S-RATEX = 58.08%. The pathological speaker's voice sample evidenced the following RATEX values: J-RATEX = 46.11% and S-RATEX = 70.42%. Therefore, the healthy speaker produced a voice sample which evidenced lower jitter and shimmer as compared to the pathological speaker.

ANOMALIES -- This parameter represents the number of occurrences within a given F0 contour in which short discontinuities appeared in the voiced segments. Recall that short discontinuities are defined as sequences of 1 or 2 F0 values which fall outside the limits of acceptable input F0 values. These short discontinuities are smoothed over by the running-median filter but are not processed for excursions due to zero values. However, their presence within an F0 contour is flagged and totaled. The value of the ANOMALIES is in percent since it represents the number of occurrences as compared to the total number of possible excursions in the data (times 100). It is expected that the number of short discontinuities will rise in voice samples which are difficult to pitch track due to degree of pathology. An ANOMALIES value of 3.99% was found for the voice sample produced by the healthy speaker as compared to a value of 4.25% found for the voice sample of the pathological speaker.

The parameters discussed so far are based on measurements of excursions from smoothed contours of F0 and A0 values produced by the non-linear smoother. There is one other perturbation parameter which measures adjacent F0 and A0 values within each respective contour to produce C2C estimates of jitter and shimmer.

J-DPF/S-DPF -- This parameter is the Directional Perturbation Factor

(DPF) which has been adapted from the work of Hecker and Kreul (1971 -- see Section 4.1.1 above). DPF is based on changes of direction in values within F0 and A0 contours. DPF counts the number of times there is a change in algebraic sign when a difference is measured between adjacent input contour values. Therefore, roughness in an F0 or A0 contour may be evaluated as small directional changes in the contour. DPF totals the number of directional changes and this total is divided by the total number of possible directional changes within a given contour and multipled by 100 to produce percentage values of J-DPF and S-DPF. The DPF parameter has been modified by including the requirement that a magnitude difference of greater than 3% must occur for any given directional change before it is included in the total DPF count. This threshold is another attempt to exclude the normal distribution of F0 directional changes from the perturbation parameters. It should be noted that the limits for acceptable F0 values within a given contour also effects this perturbation parameter. Any zero in the F0 contour is not evaluated by the DPF algorithm and this includes onsets/offsets of voicing as well as short discontinuities within the contour.

Figure 5.16 presents segments of unsmoothed F0 values (5.16a) and unsmoothed A0 values (5.16b) derived from the pathological speaker's data presented in Figs. 5.11a and 5.12a respectively. In each figure, 20 samples of the F0 and A0 contours are displayed. The unsmoothed output of the parallel processor is shown in these figures to demonstrate the measurement of DPF for F0 and A0 contours. In Figure 5.16a, a small segment of the F0 contour is displayed with stars (*) marking the values of the F0 samples. Note that within the contour there are a number of occurrences when the

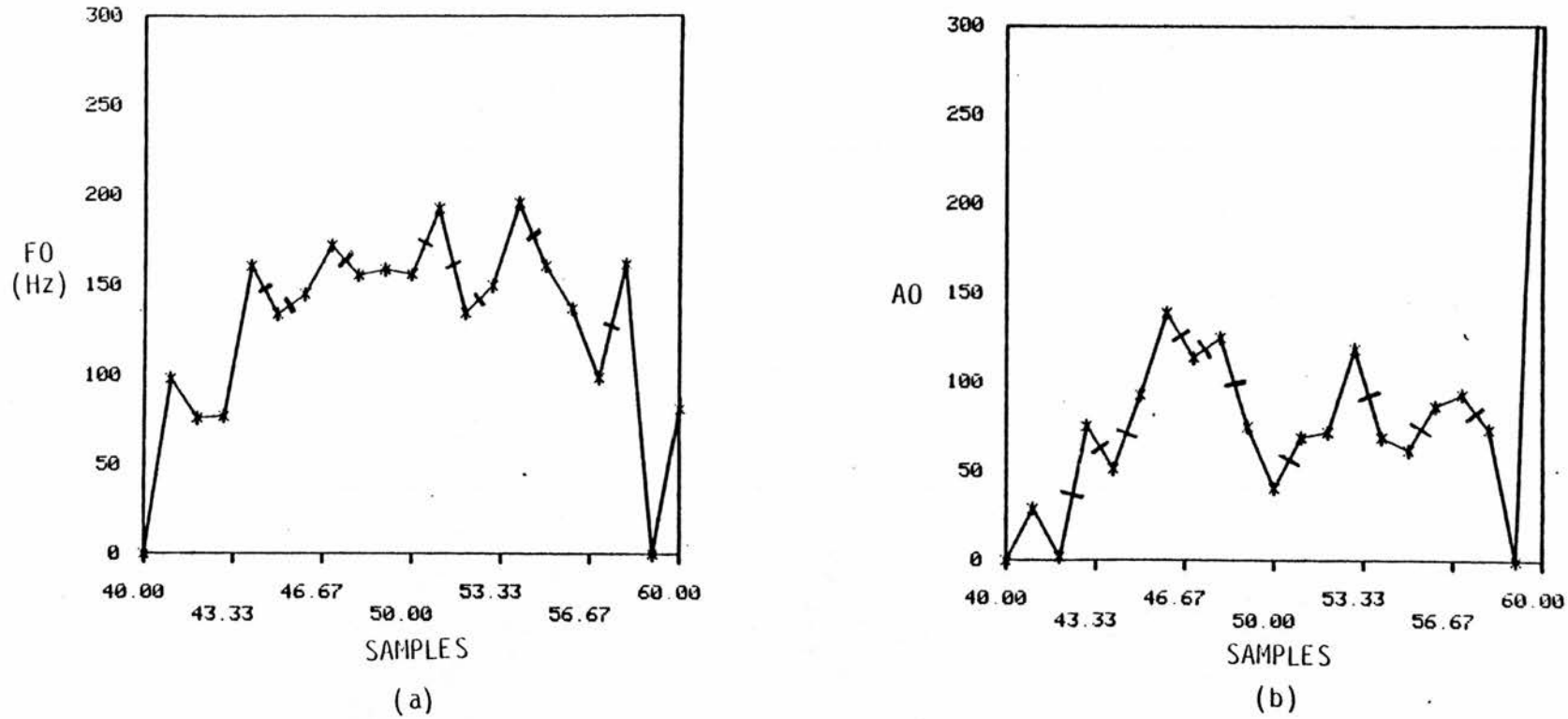


Figure 5.16 Two examples of measuring directional perturbations from the unsmoothed F0 (a) and A0 (b) contours extracted from a sample of connected speech produced by the male pathological speaker. The slashes indicate those changes in direction with magnitudes of greater than 3%.

value of F0 changes direction from its previous two values. The directional changes which exceed a 3% change in magnitude are denoted by hashes across the connecting solid lines. In this example, there are 9 significant directional changes which indicate a fairly perturbed section of the overall F0 contour. Figure 5.16b is the equivalent A0 contour (greatly enlarged for this demonstration) of the F0 contour in Fig. 5.16a. The directional changes with significant magnitudes are marked with hashes in this Figure as well. In fact, all 10 possible directional changes are significant for this segment of the overall A0 contour which suggests a very perturbed waveform. Note that in both Figs. 5.16a and b that the zero values do not contribute to the measurement of directional changes. The DPF analysis for the voice sample produced by the healthy speaker demonstrated a J-DPF of 15.1% and S-DPF of 26.2% while for the pathological speaker, a J-DPF of 30.3% and S-DPF of 40.6% was found. It would appear that the DPF measures of the pathological speaker are substantially greater than the parameters found for the healthy speaker.

SECTION 5.5 -- OUTLINE OF ANALYSIS PROCEDURES FOR EVALUATING WAVEFORM PERTURBATIONS

The following outline summarizes the procedures used to analyze a voice sample for long-term parameters of intonation and perturbation.

1. A recorded voice sample is low-pass filtered by an anti-aliasing filter (stopband frequency = 10 KHz) and digitized at a sampling rate of 20 KHz. The digitized data is filed for further analysis.
2. The sampled data is analyzed by interactive means to determine

the peak amplitude of the background noise. The peak amplitude is used in the threshold system for determining the presence of silence within the voice sample. It is also determined at this time whether the signal requires inversion in order to bring the prominent peaks of the waveform in to the positive domain.

3. Phase compensation of the voice sample is performed using a compensation factor specified for the given tape recording/playback system. The phase compensation is completed using an analysis interval, frame rate and FFT size equal to 409.6 ms (8192 samples at a 20 KHz sampling rate). See Chapter 6 for a complete discussion of the phase compensation procedures.

4. The phase-compensated voice sample is low-pass filtered with a 32-pole linear phase filter -- the filter cutoff is set to 600 Hz for males and 800 Hz for female speakers.

5. F0 and A0 contours are extracted from the low-pass filtered voice sample using the parallel processor (the analysis interval is set to 25 ms for male speakers and 20 ms for females). This step is completed two times. Firstly, pitch detection is completed with a sex-specific setting of the analysis interval shift rate (10 ms for males and 5 ms for females) to determine the median pitch period for a given voice sample. Then, the pitch analysis is repeated with the shift rate equal to the median pitch period as determined from the first pass.

6. The non-linear smoother consisting of a 5-point running-median and a 3-point Hanning window is applied to the unsmoothed F0 and A0 contours to determine smoothed trend lines and excursions. The RATEX and DPF parameters of jitter and shimmer are also calculated at this time.

7. Distributional statistical analyses of fundamental frequency,

signed and magnitude excursions in percent are completed.

8. All intonational and perturbation parameters are stored in the voice acoustics filing system (VAFS) as a permanent record of the perturbation analysis for a given speaker.

CHAPTER 6

THE APPLICATION OF THE PERTURBATION MEASUREMENT SYSTEM TO SPEAKERS EVIDENCING HEALTHY AND PATHOLOGICAL VOICE CONDITIONS

CHAPTER 6

THE APPLICATION OF THE PERTURBATION MEASUREMENT SYSTEM TO SPEAKERS EVIDENCING HEALTHY AND PATHOLOGICAL VOICE CONDITIONS

6.0 INTRODUCTION

In Chapters 3 and 5, a number of algorithms were described for a system which extracts a variety of acoustic parameters of intonation and perturbation from recorded samples of connected speech. The main aim of constructing such a system is its eventual use as a tool for screening the population for speakers who evidence pathological conditions of the voice. The experiments to be presented in this chapter are applications of the perturbation measurement system to voice samples recorded from groups of pathological and healthy speakers -- these studies are the first steps on the way to fulfilling the aim of providing a system capable of screening populations for voice pathologies.

The experiments are divided into two broad types of investigation. The first 2 investigations are concerned with the required nature of the connected speech samples which are to be input to the perturbation measurement system. In Section 6.1, an experiment is presented which determines the durational length requirement of a given sample of connected speech in order that long-term intonational and perturbational parameters are captured which characterize a speaker's phonatory efficiency. The effects of compensating for low-frequency phase distortion associated with the tape recording of connected speech samples is investigated in Section 6.2. The second type of experiments is concerned with the

differentiation between pathological and healthy speakers -- a series of parametric statistical procedures are applied to the acoustic parameters produced by the analysis of recorded voice samples by the perturbation measurement system. Procedures are described in Section 6.3 for the selection and evaluation of the control and pathological groups of speakers to be used in the remaining statistical analyses. In section 6.4, the effects of both voice condition and speaker gender on the differentiation of speakers by the acoustic parameters are investigated by analysis of variance. The types and degrees of first order correlations between the various acoustic parameters are presented for each of the experimental speaker groups in Section 6.5. Finally, in Section 6.6, pattern recognition experiments are completed in which pathological and control speakers are classified by the Maximum Likelihood principle. The classification tasks will determine the best acoustic parameters for detecting pathological and control speakers as well as establishing the actual success of detection based on these features.

SECTION 6.1 -- DURATIONAL REQUIREMENTS OF THE CONNECTED SPEECH SAMPLE FOR PERTURBATION ANALYSIS

An important consideration in many applications of automatic speaker characterization is establishing the minimum sample duration for long-term acoustic parameters to reach stability. Stability is here understood to refer to the extraction of parameters in such a fashion that they genuinely characterize the speaker rather than the message content of the speech sample. The characterization of a speaker by long-term acoustic parameters is the first step in a

variety of speech pattern recognition experiments. For example, certain types of speaker recognition studies are designed to be text-independent with the major restriction on the stimulus materials being their overall length (Rosenberg 1976). Similarly, the description of characteristic voice quality in healthy and pathological speakers requires a speech sample long enough to permit the abstraction of long-term average features of overall quality from the fluctuating values of short-term segmental performance (Laver et al. 1981).

The long-term behavior of fundamental frequency has been shown to be an easily available acoustic parameter for characterizing speakers (Atal 1976). Most studies of fundamental frequency in this area are based on long-term feature averaging. Fundamental frequency is one of a variety of speech parameters for which Markel, Oshika and Gray (1977) suggest that the concept of a long-term average value is relevant, despite the lack of either a true mean or true variance in F0 data in real speech due to non-random factors such as the declination effect in intonation. One major aspect of long-term features of F0 is the overall intonational behavior of a given speech sample -- included in this category are such detailed measures as mean, median, mode, standard deviation, skewness, kurtosis, etc. For long-term intonational F0 statistics, Nolan (1983) noted a general finding within the relevant literature that within-speaker variation was minimized for durations of approximately one minute. Steffan-Batog, Jassem and Gruzska-Koscielak (1970) found that 50 seconds of read text was enough to produce a regular distribution for long-term averaging of F0. Green (1972) suggested that segments of speech as short as 15

seconds would flatten the short-term variations of pitch in samples of conversational speech. Distributional convergence of F0 statistics occurred for a sample duration of approximately 60 seconds for read scripts analyzed by Horii (1975). Mead (1974) reported that an optimal duration of 75 seconds was useful for a speaker recognition task using unconstrained speech, but durations as short as 30 seconds approximated longer duration results in cases where only short samples of speech were available for analysis. Markel and Davis (1979) noted that durations between 40 and 70 seconds were sufficient to show convergence to a stable long-term average F0 for linguistically-unconstrained speech.

The comments above concern intonational behavior, where the frequency value of the general trend line through the F0 data is of greater relevance than very local irregularities in adjacent periods. But of course, on close inspection, the succession of pitch periods making up the fundamental frequency contour of voiced speech does not show a perfectly smoothly-changing sequence of duration values. The duration of each successive pitch period tends to vary randomly from the general intonational trend line discernible through a sequence of pitch periods. These local deviations of individual periods from the smooth intonation are considered perturbations of the F0 contour, as explained earlier, and auditorily perceived as a 'rough' phonatory quality (Hiller et al. 1983). Such perturbatory deviations are usually larger in degree and more frequent in speakers suffering from laryngeal pathology than in healthy speakers. Because of their role as signals of potential laryngeal disorder, therefore, short-term perturbatory movements of F0 have often been measured on a long-term

statistical basis for the assessment of such voice disorders (see especially, for example, Lieberman 1963; Hecker and Kreul 1971; Davis 1976; Askenfelt and Hammarberg 1980; 1981; Laver et al. 1982, and Chapter 4 above of this thesis in general). The long-term measures of perturbatory F0 behavior typically include the mean, range and rate of occurrence of these perturbations in voiced speech. In most studies of perturbation, F0 acoustic parameters are derived from sustained vowel phonations or short sentences. It was noted in Chapter 4 above that only a few studies have attempted to extract data from longer durations of continuous speech, which can be considered a more natural use of phonation (Askenfelt and Hammarberg 1980; 1981; Laver et al. 1982).

The literature reported immediately above suggests that intonational values stabilize to a steady mean in speech data whose minimum duration lies somewhere between 40 and 75 seconds. But a corresponding minimum duration for stabilization of perturbational values is not yet well established, and the purpose of this study is to determine the minimum duration of continuous read speech which will stabilize long-term acoustic parameters of perturbation in healthy speakers. The opportunity will also be taken to further the study of the minimum duration for stabilization of intonational characteristics.

SECTION 6.1.1 -- SPEAKERS, SPEECH MATERIAL AND ANALYSIS PROCEDURES

High-quality tape recordings were made during oral reading of the first two paragraphs of "The Rainbow Passage" (Fairbanks 1960) by 20 adults (10 males and 10 females) who had no known history of speech or voice disorders. Nineteen of the speakers were British

(Scottish or English) with the remaining speaker being an American. Smokers were not excluded from this investigation. Prior to the recording, each speaker familiarized himself with the passage and was asked to read at a comfortable loudness level. The recorded speech samples were digitized using a PDP 11/40 computer and further processing of the data was completed on a VAX 11/750 computer. For subsequent analyses, the total utterance of each speaker was divided into 5 sec segments. Intonational and perturbatory statistics were completed for durations successively incremented by 5 seconds (i.e. 5, 10, 15 ,..., 60 seconds) up to 60 seconds or the end of the utterance, whichever came first. Each incremental analysis therefore incorporated information from the previous shorter duration analyses.

The measurement system used in this study for deriving long-term F0 and perturbation parameters is essentially the one described in Chapters 3 and 5. However, there are some notable differences in the analysis procedures used in this particular study. Firstly, the voice samples recorded from the 10 male speakers were digitized at a sampling rate of 10 KHz. It was during the completion of this study that the concern arose over the intrinsic accuracy of F0 results derived from speech signals digitized at 10 KHz. As discussed in Section 3.3 above, a sampling rate is required which provides a reasonable compromise between measurement accuracy associated with the production and perception aspects of speech. To reach this compromise in measurement accuracy, a sampling rate of 20 KHz would be suitable for quantizing the speaker-characterizing perturbations evidenced in voices samples recorded from female speakers and children. Following the analysis

of the 10 male speakers, it was decided that the sampling rate should be increased to 20 KHz for all voice samples. Thus, the voice samples recorded from the 10 female speakers were sampled at 20 KHz for this study. Secondly, phase compensation techniques were not applied to any of the 20 voice samples analyzed for this study. Pre-processing of each voice sample by low-pass filtering was completed prior to digitization by an analog filter which also prevented aliasing of the input signal. Thirdly, the set of parameters analyzed for long-term behavior consisted only of 6 features including 1) the intonational parameters F0-AV and F0-DEV and 2) the frequency perturbation parameters J-AVEX, J-DEVEX, J-RATEX and J-DPF. Analyses of the voice samples for amplitude perturbation parameters were not completed for this study.

SECTION 6.1.2 -- RESULTS AND DISCUSSION -- MALE SPEAKERS

Figures 6.1 to 6.6 display examples of the results of the present study for 6 acoustic parameters measured as a function of increasing sample duration (incremented every five seconds) for the 10 male speakers. The six parameters are the F0-AV in units of Hz (Figure 6.1), the F0-DEV in units of Hz (Figure 6.2), the J-AVEX in percent (Figure 6.3), the J-DEVEX in percent (Figure 6.4), the J-RATEX in percent (Figure 6.5) and the J-DPF in percent (Figure 6.6). The four perturbation parameters are measured in percent to normalize for differing levels of F0. The results for each of the 10 male speakers are labeled in each Figure as A - J, with each Figure representing the parameter on the ordinate versus increasing duration in seconds on the abscissa. It should be recalled that each 5 sec increment on the abscissa includes data from the previous

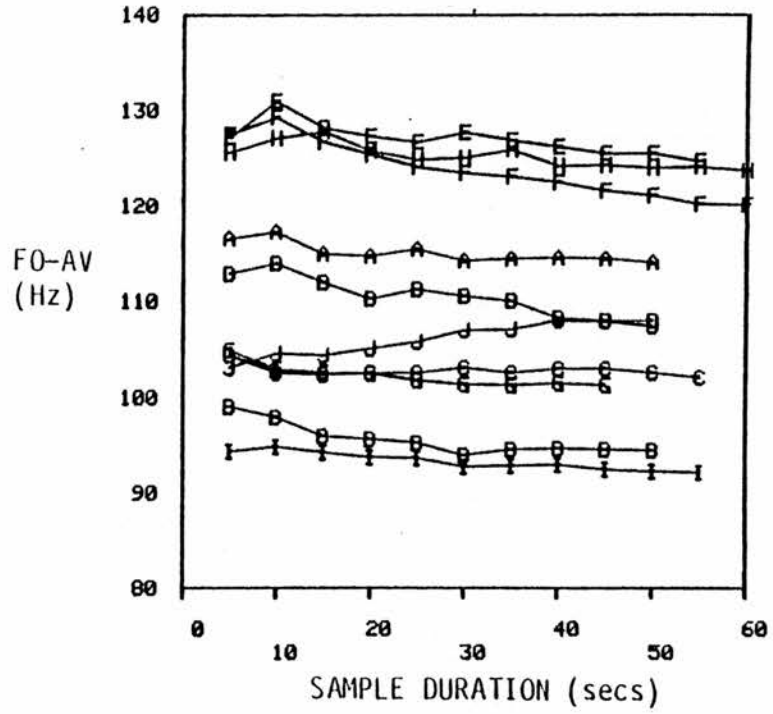


Figure 6.1 Changes in long-term value of FO-AV with increasing sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

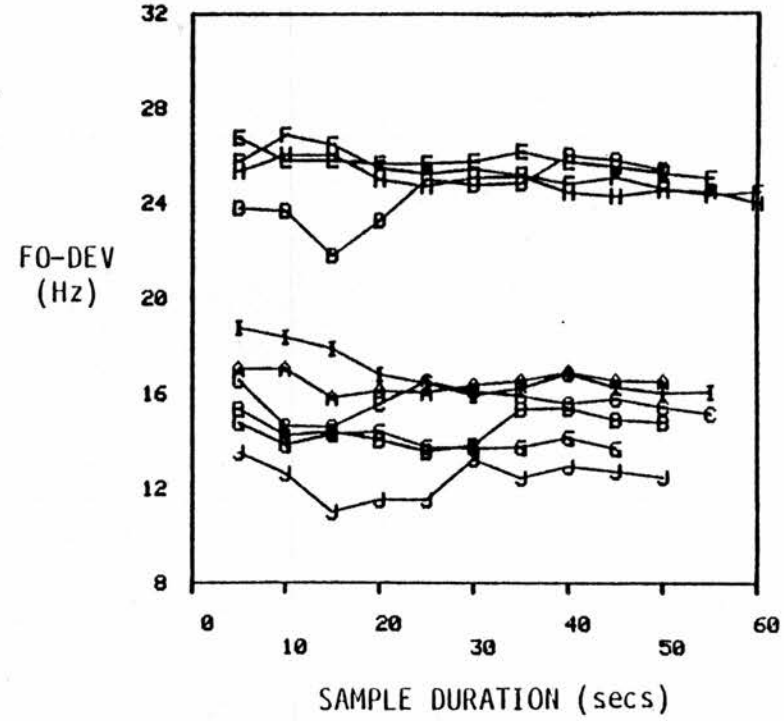


Figure 6.2 Changes in long-term value of FO-DEV with increasing sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

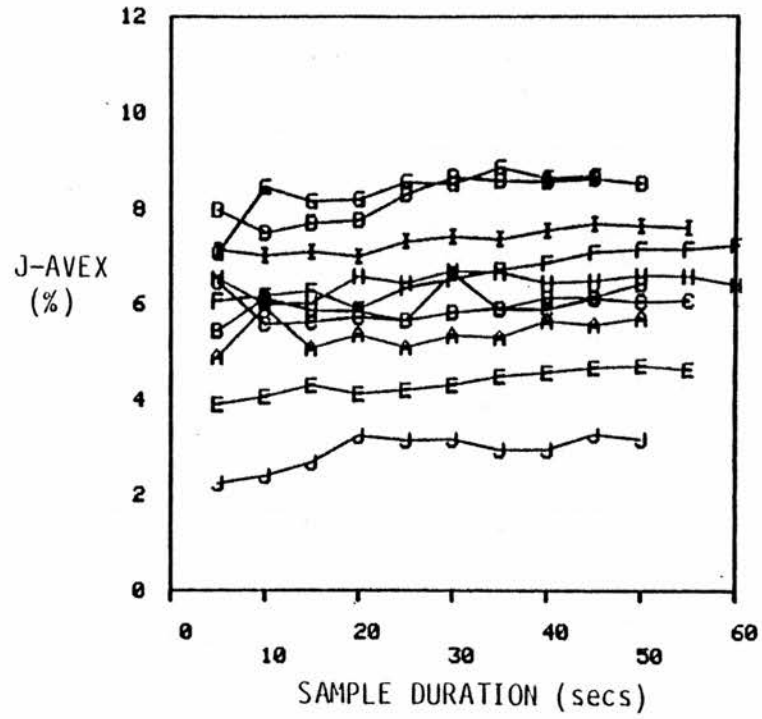


Figure 6.3 Changes in long-term value of J-AVEX with increasing sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

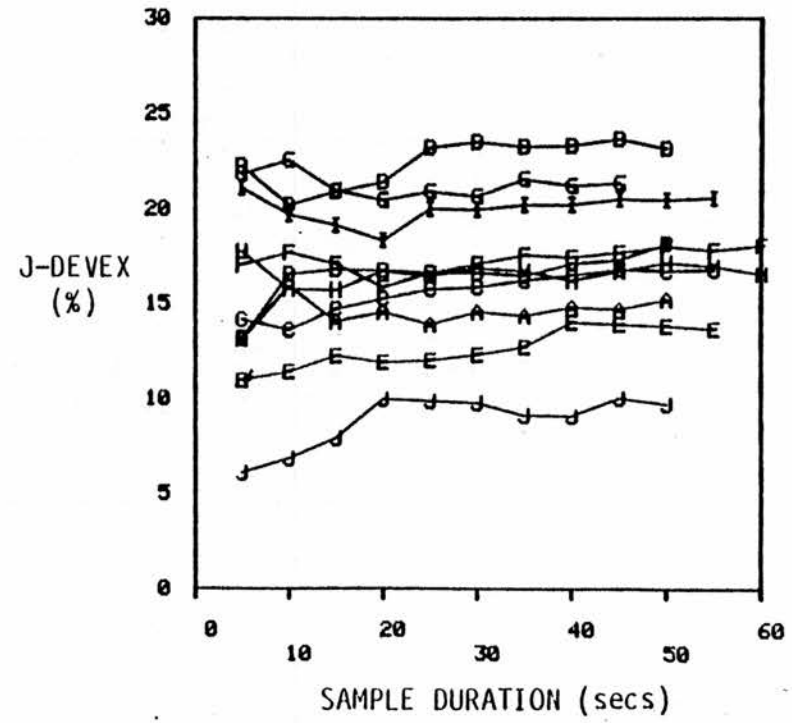


Figure 6.4 Changes in long-term value of J-DEVEX with increasing sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

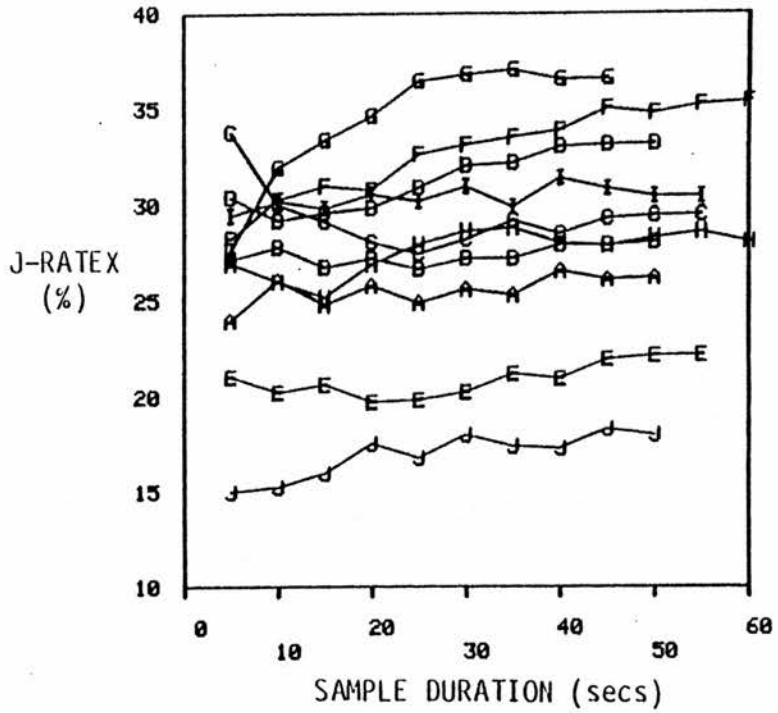


Figure 6.5 Changes in long-term value of J-RATEX with increasing sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

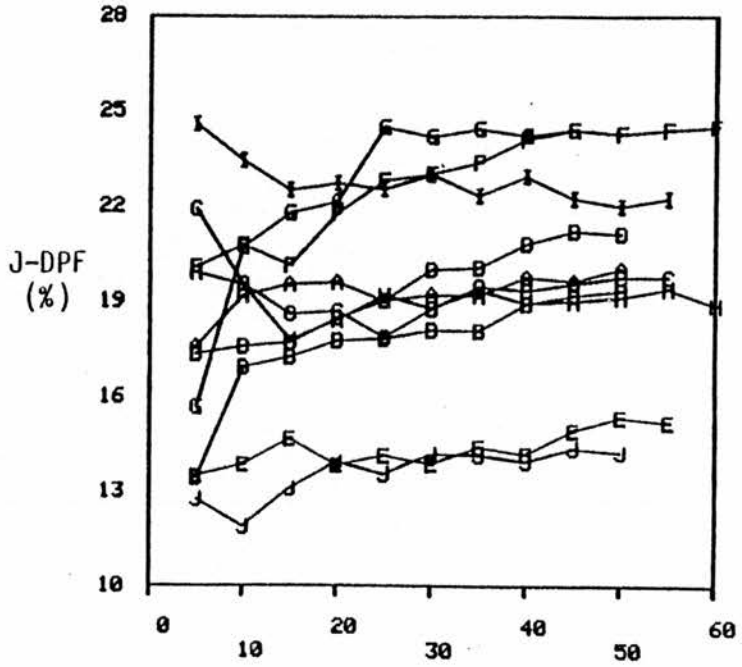


Figure 6.6 Changes in long-term value of J-DPF with increasing sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

segments up to and including the specified duration. In addition, each duration analyzed for each speaker represents the actual length of the recorded speech up to a maximum of 60 seconds, which includes both voiced and unvoiced speech. It was found that the ratio of voiced to unvoiced data points increased linearly as duration was increased for all the male speakers' data. This finding suggests that the male speakers spoke the text with a regular tempo and no intermittent long periods of voicelessness. The average ratio of voiced to unvoiced data points was approximately 66% for male speakers and 70% for females.

The general description for the 10 speakers' durational curves for all 6 parameters is one of instability at the shorter durations below approximately 25 seconds followed by a flattening of each parameter by the end of the spoken text. Visual estimates of the data suggest that 40 seconds of spoken text is a sufficient duration for all 6 parameters to reach stability. A value of 40 seconds of spoken text agrees broadly with the findings of other researchers, particularly those studies in which a standard text was used as the stimulus material. No large differences were noted between intonational and perturbational measures for overall duration required for parametric stability (comparing, for example, Figures 6.1 and 6.3 which display the F0-AV and J-AVEX, respectively). The general growth patterns for the intonational and perturbational parameters were dissimilar -- the intonational measures demonstrating values which decreased with increasing duration while the perturbation measures display increasing values with time. For example, in Figure 6.1, the F0-AV is seen to lower slightly in frequency as more data is accumulated for most of the speakers. The

decrease in F0-AV may be the result of 2 effects: 1) paragraph effects associated with the linguistic structure of read English and 2) (possibly) progressively decreasing tension (i.e. decreased stiffness) of the vocal folds during continuing oral reading. The notable exception to the downward trend for F0-AV is male speaker J, who demonstrates increased F0-AV with increased duration of the data. The F0-AV values for the 10 speakers are spread across a frequency range reported by other researchers for healthy male voices. The downward trends of Figure 6.1 are not quite so clear in Figure 6.2 for the F0-DEV parameter. Though the overall trend appears to be decreasing F0-DEV values as sample duration increases, a number of male speakers (in particular, speakers A, B, D and J) produced increasing F0-DEV values with increasing sample duration. The perturbation measures displayed in Figures 6.3 to 6.6 demonstrate the tendency towards increased perturbation values with time which may be correlated with decreased efficiency of phonation with progressive laryngeal fatigue experienced in the speaking of long texts. These growth curves for perturbation measures may provide some useful information about voice pathology, following the precedent of the notion of articulation growth curves for speech discrimination testing in audiology.

SECTION 6.1.3 -- RESULTS AND DISCUSSION -- FEMALE SPEAKERS

Figures 6.7 to 6.12 present the results for 10 female speakers for the same 6 parameters measured as functions of incremented sample duration. Note that the vertical scales of Figures 6.7 and 6.8 differ from Figures 6.1 and 6.2 in order to accommodate the female speakers' F0 values. In each figure, the parameter value is

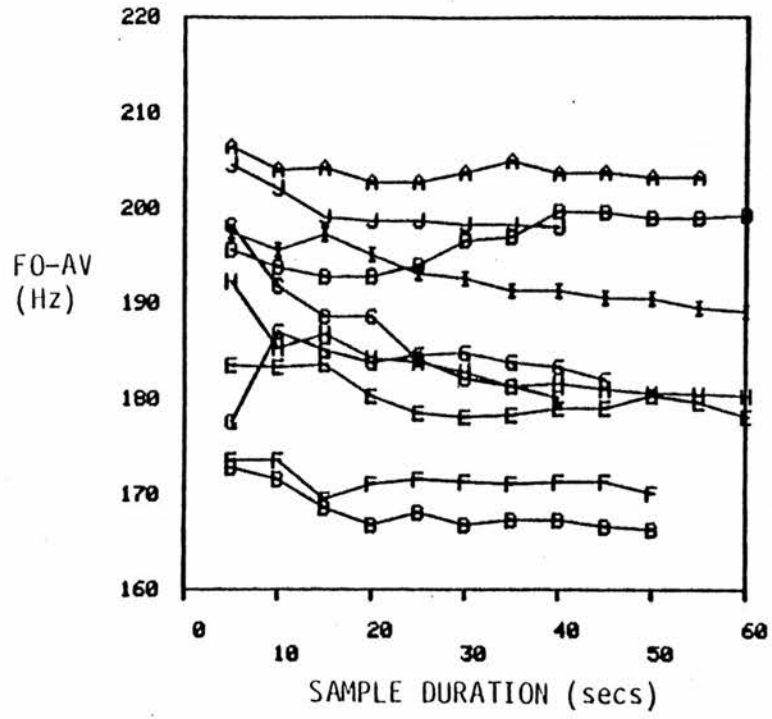


Figure 6.7 Changes in long-term value of FO-AV with increasing sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

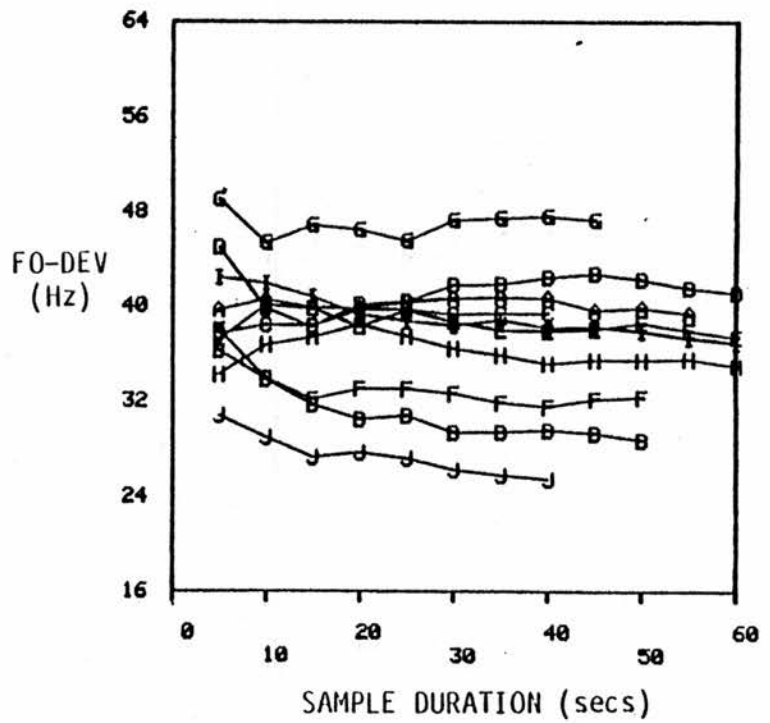


Figure 6.8 Changes in long-term value of FO-DEV with increasing sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

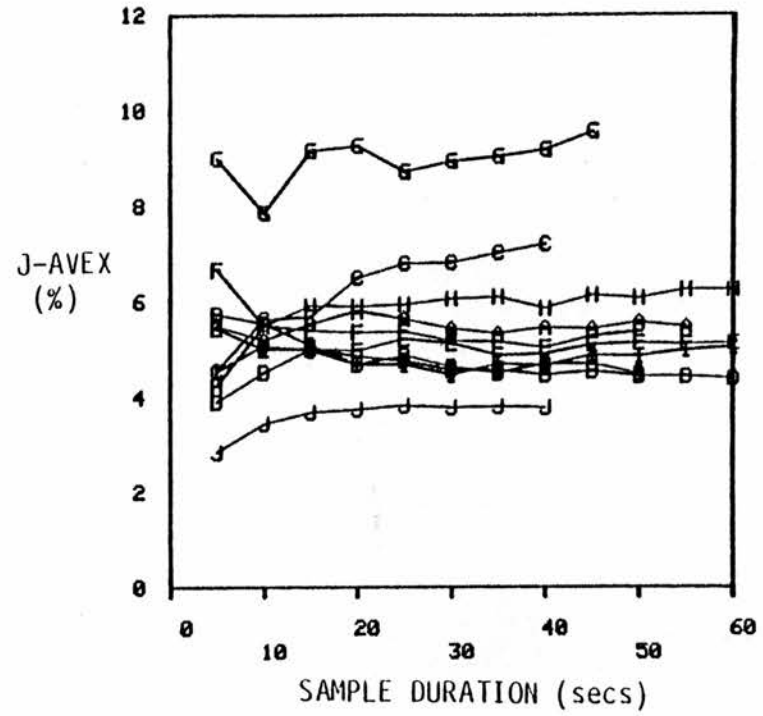


Figure 6.9 Changes in long-term value of J-AVEX with increasing sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

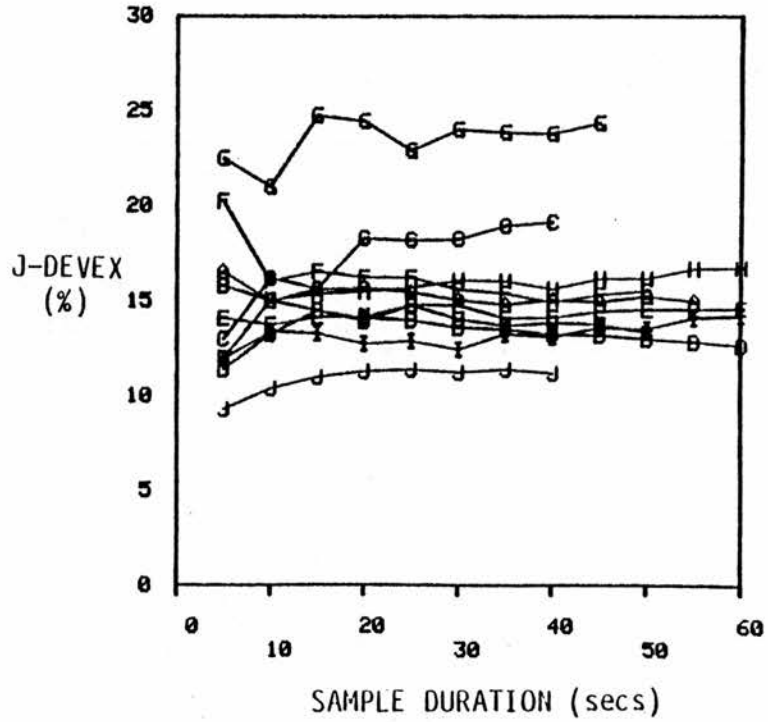


Figure 6.10 Changes in long-term value of J-DEVEX with increasing sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

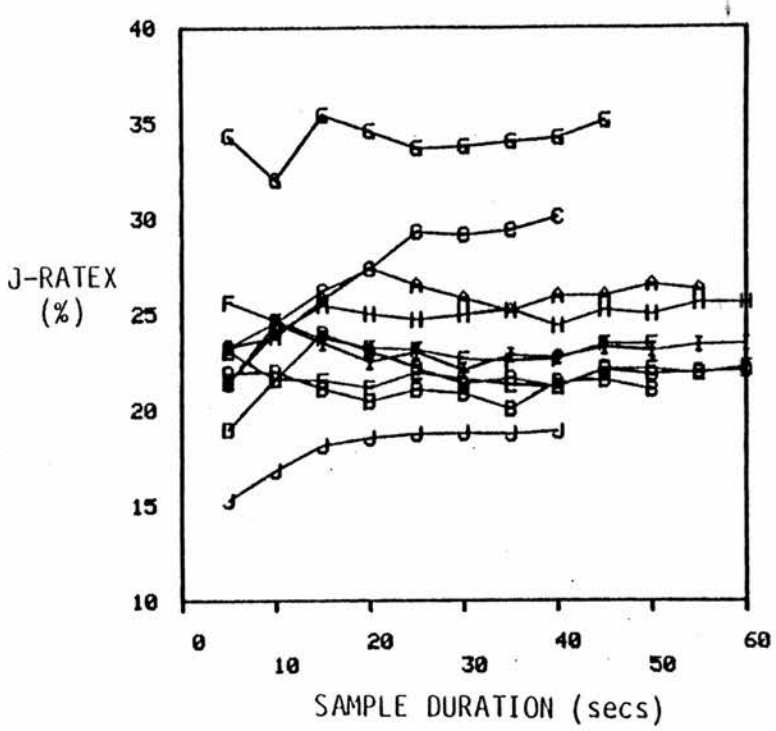


Figure 6.11 Changes in long-term value of J-RATEX with increasing sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

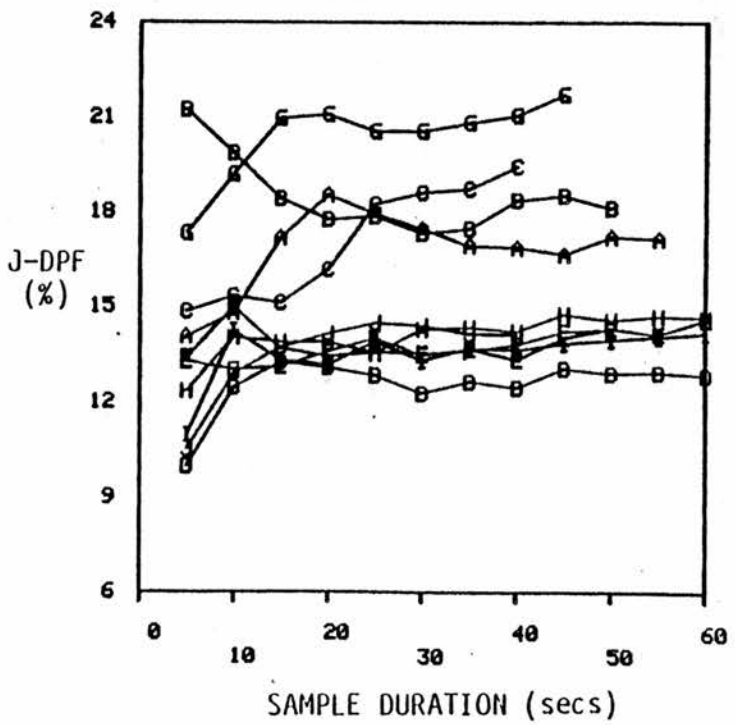


Figure 6.12 Changes in long-term value of J-DPF with increasing sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

displayed versus sample duration, and the 10 female speakers labeled A - J. The growth patterns of the parameters derived from the female data are similar in nature to the male results. Early instability of the parameter values is followed by movement towards stability. The long-term duration of 40 seconds for relative parametric stability also applies to the female results. As seen in Figure 6.7, growth curves for the F0-AV parameter demonstrated decreasing values with increasing sample duration though female speaker D produced increased F0-AV values with increased duration. As was the case for the male speakers, the overall trends for the F0-DEV values derived from the female speakers' voice samples (Figure 6.8) are not quite as clear as seen for the F0-AV parameter (Figure 6.7). There is a tendency towards a general downward trend of F0-DEV as sample duration increases but a number of speakers (in particular, speakers C, D and G) demonstrate increasing F0-DEV values. The overall values of the F0-AV parameter for the female speakers are nearly an octave greater than the F0-AV values produced by the male speakers. The growth curves for the perturbation parameters (as seen in Figs. 6.9-12) revealed increased values as duration was increased for the female speakers, but not with the same consistency as found for the male speakers' curves. Indeed, a number of the growth curves for the perturbation parameters derived from the female speakers' voice samples appear to vacillate in a manner which does not support a strong trend towards increasing or decreasing parametric values as sample duration increases. The resultant perturbation values for the female speakers are similar to the male results which suggests that the perturbation analysis method successfully normalizes for differing levels of F0. One notable result was female speaker G who demonstrated perturbation

values which are much higher than the other speakers. This speaker's results may be related to her history of heavy smoking. For all the parameters, the distributions of parametric values were narrower for the female speakers as compared to the male speakers.

SECTION 6.1.4 -- FURTHER ASSESSMENT OF THE DURATIONAL DATA

To assess the durational data, absolute difference curves were derived from the data contained in the original durational curves. A difference curve is composed of values which are the absolute differences between a speaker's final long-term parametric value for a read passage and each cumulative value at each 5 sec time increment. Figures 6.13 to 6.18 display the absolute difference curves for the 6 parameters derived from the voice samples of the 10 male speakers. The corresponding difference curves for the female speakers are presented in Figs. 6.19-24. The vertical axis of each figure represents the difference (in Hz or percent) for each cumulative value at a given time increment (the abscissa) from the final long-term parametric value. It should be noted that the vertical scales in Figures 6.13 and 6.14 (the male speakers) differ in magnitude from Figs. 6.18 and 6.19 (the female speakers) due to the larger difference measures of some of the female speakers. Figures 6.13 and 6.18 display the difference curves for the F0-AV parameter which were derived from the male and female growth curves, respectively. The speakers are labeled A-J in each figure and correspond to the speaker labels presented in the previous figures. The difference curves in Figs. 6.13 and 6.18 demonstrate decreasing differences in Hz as the time increments approach the final durational values for F0-AV (which naturally drop to zero for the

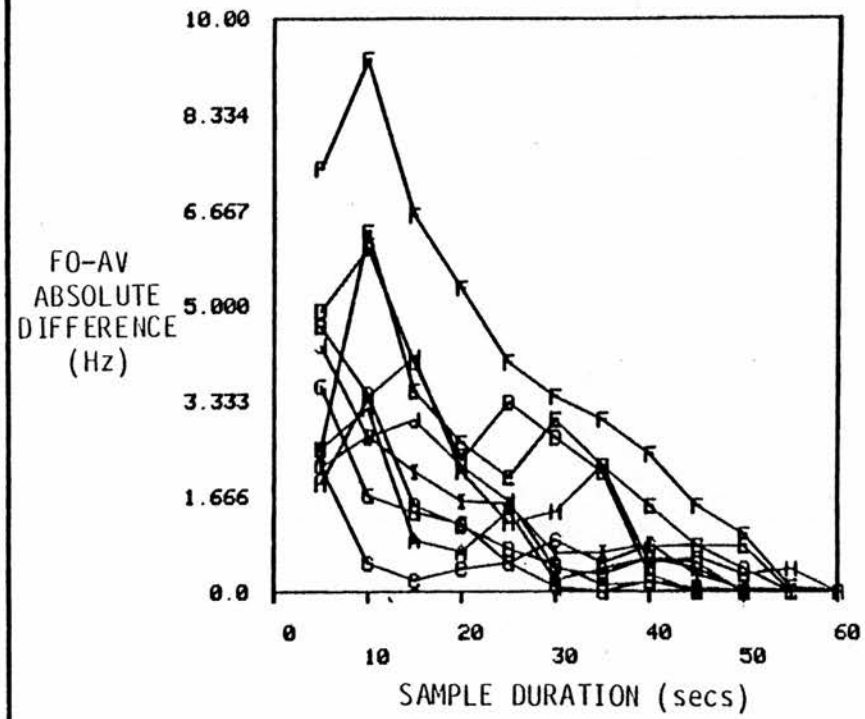


Figure 6.13 Absolute differences of FO-AV from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

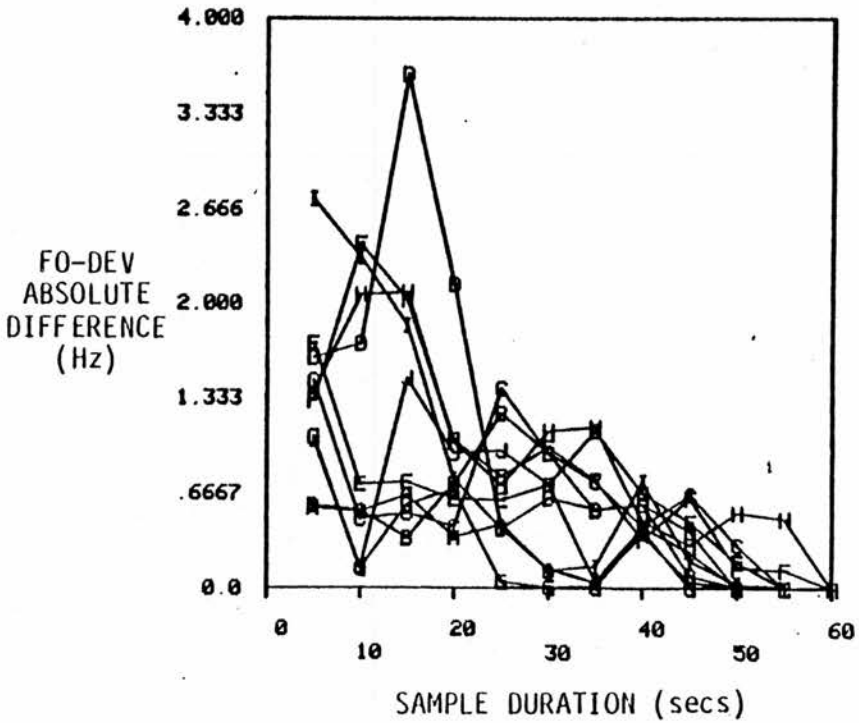


Figure 6.14 Absolute differences of FO-DEV from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

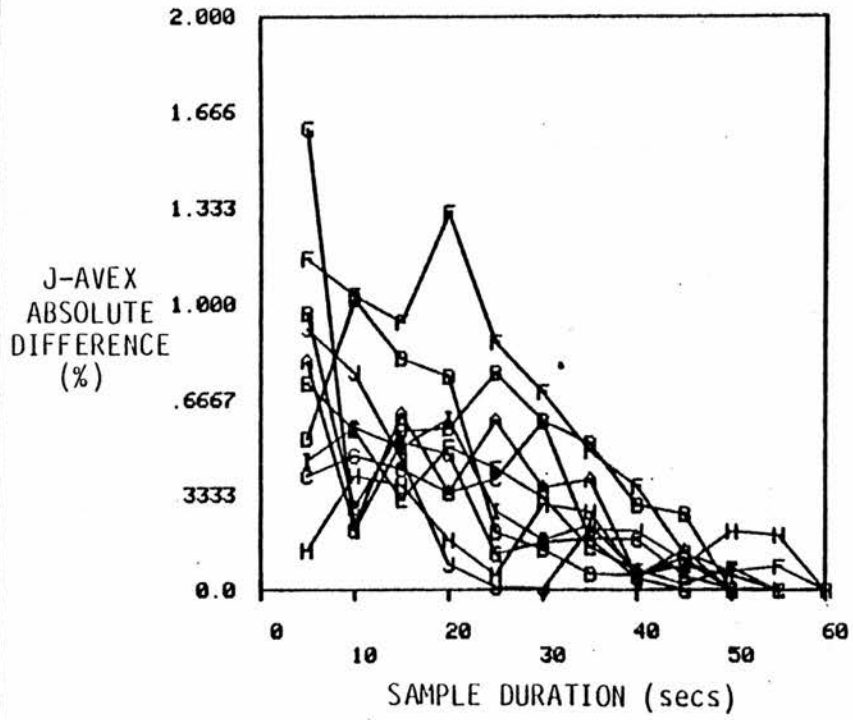


Figure 6.15 Absolute differences of J-AVEX from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

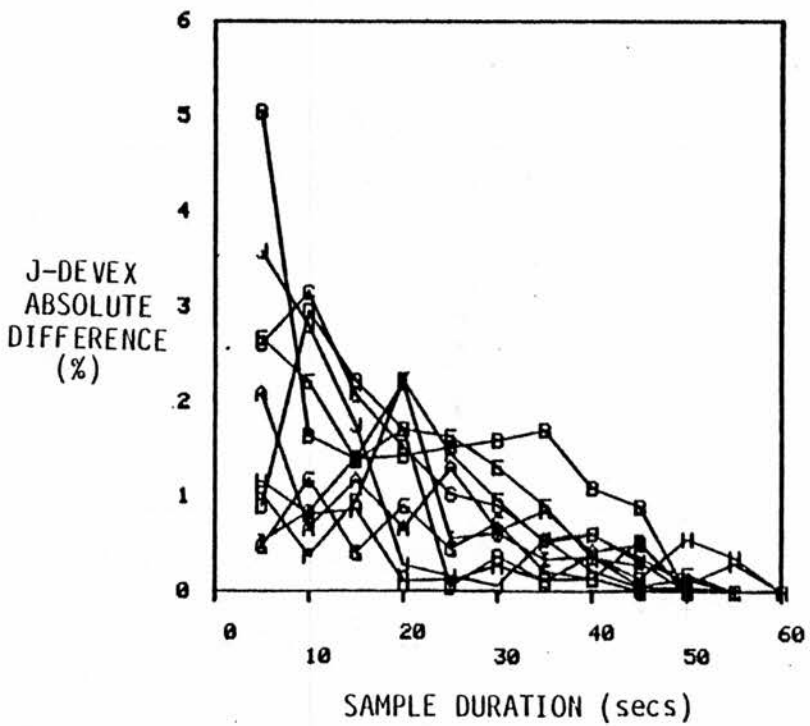


Figure 6.16 Absolute differences of J-DEVEX from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

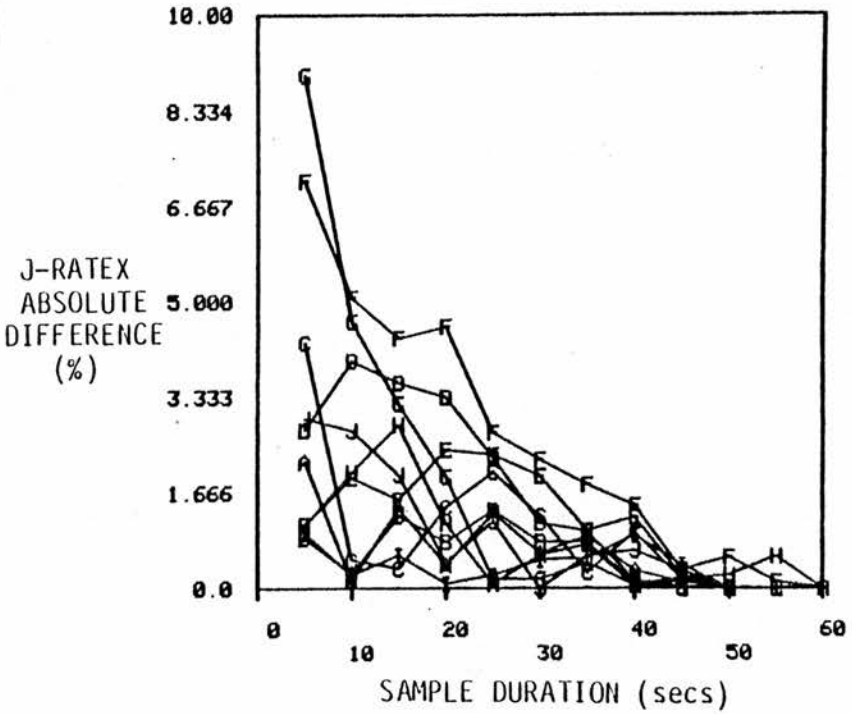


Figure 6.17 Absolute differences of J-RATEX from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

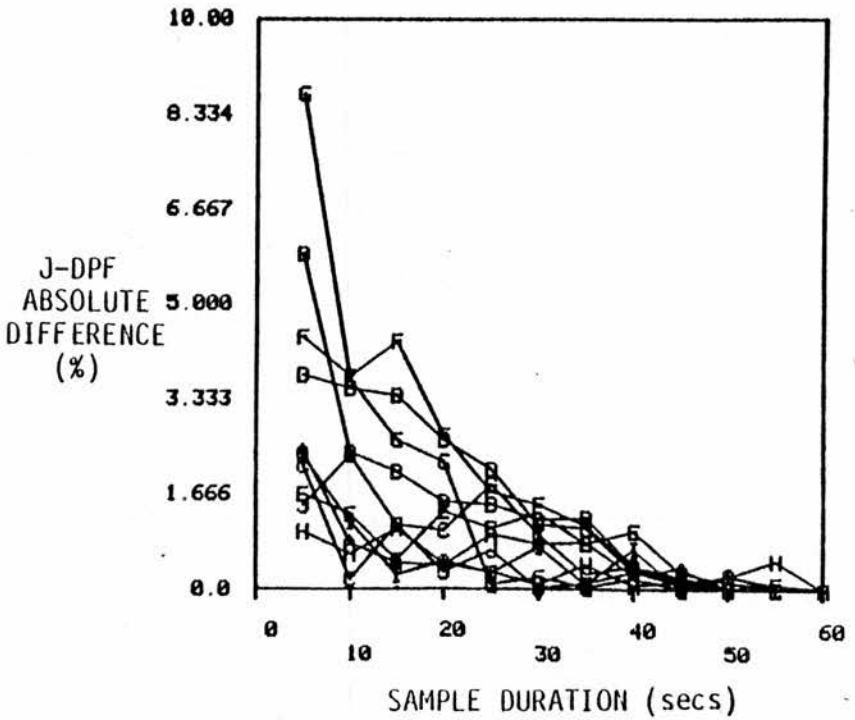


Figure 6.18 Absolute differences of J-DPF from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy male speakers (labeled A-J).

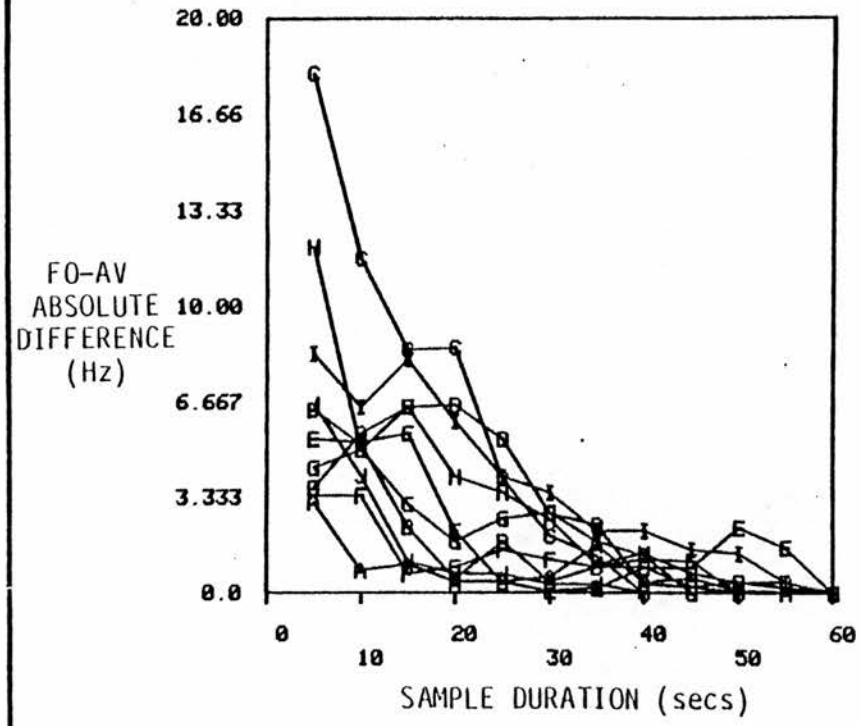


Figure 6.19 Absolute differences of FO-AV from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

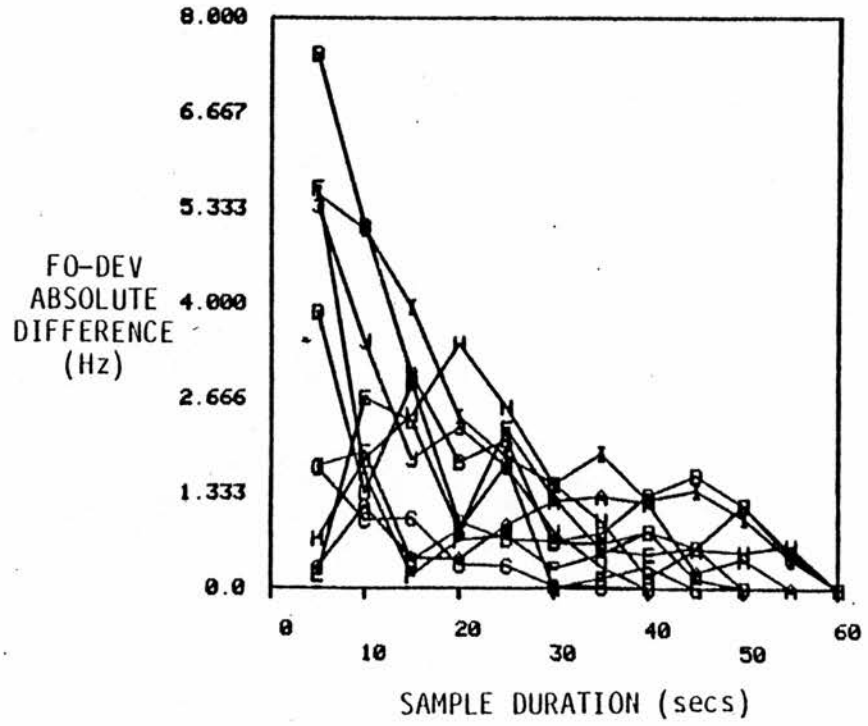


Figure 6.20 Absolute differences of FO-DEV from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

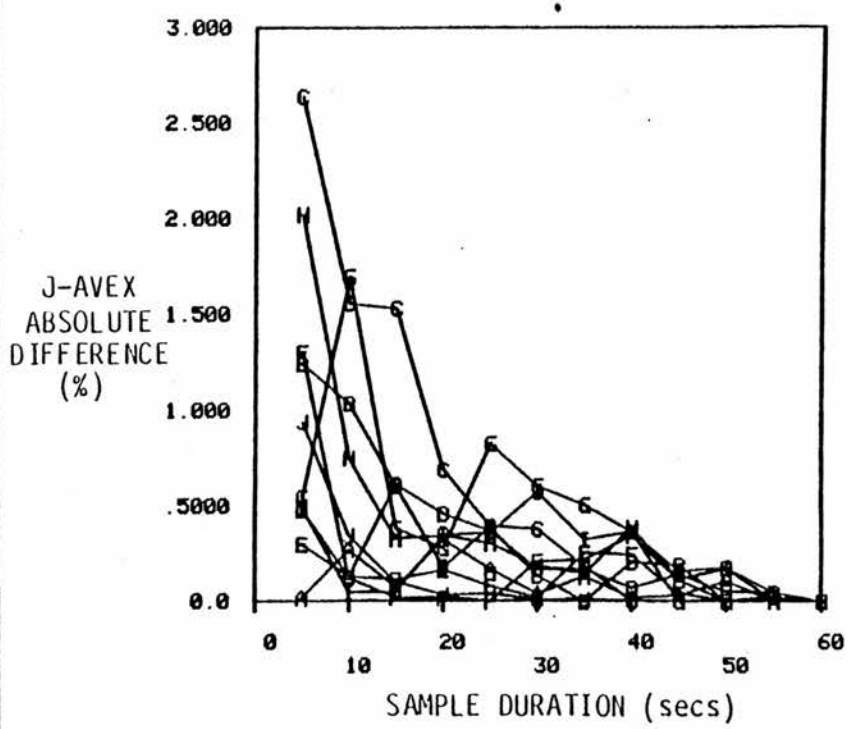


Figure 6.21 Absolute differences of J-AVEX from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

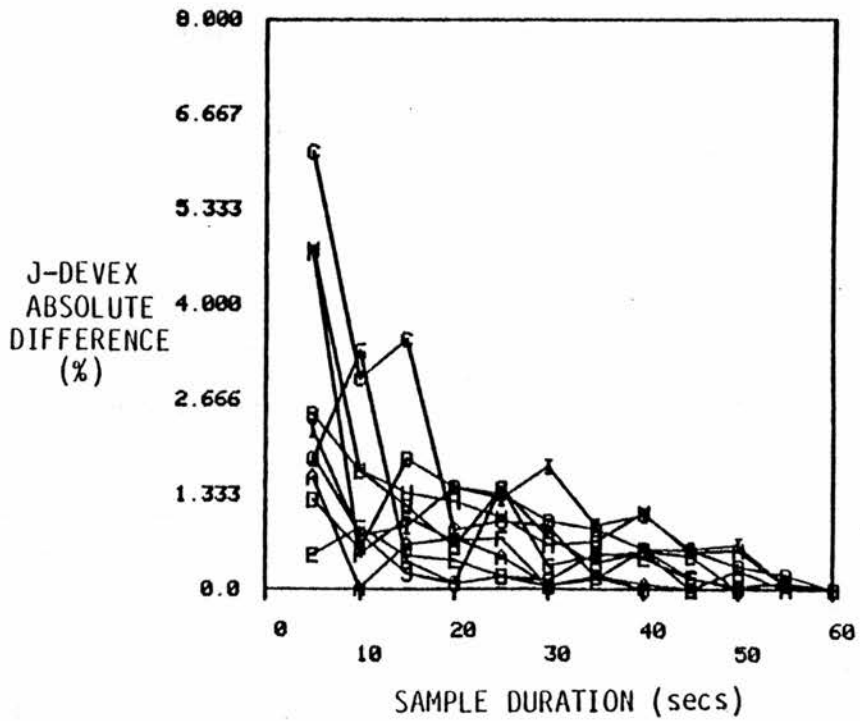


Figure 6.22 Absolute differences of J-DEVEX from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

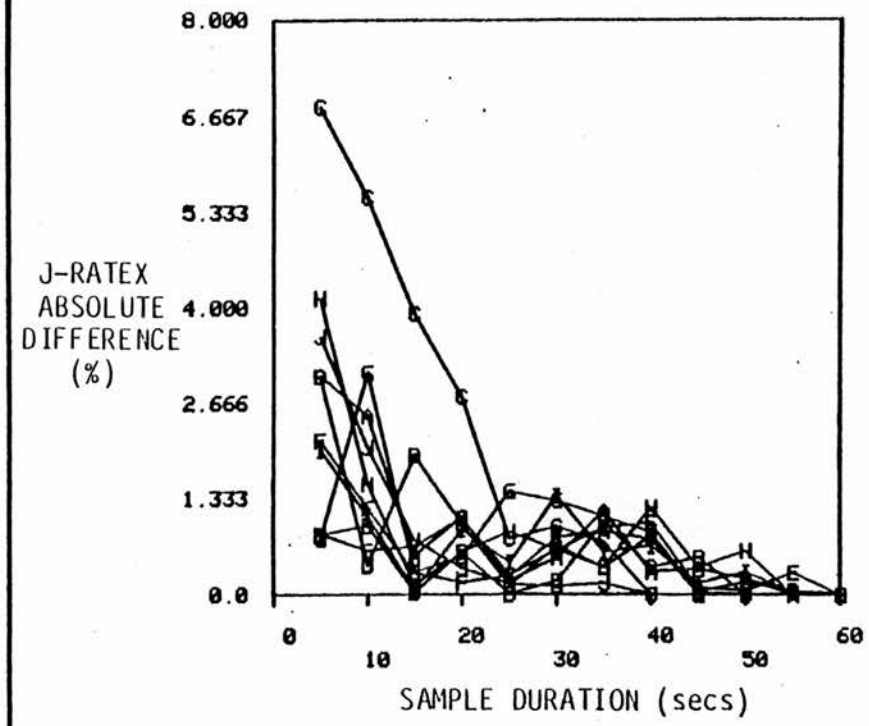


Figure 6.23 Absolute differences of J-RATEX from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

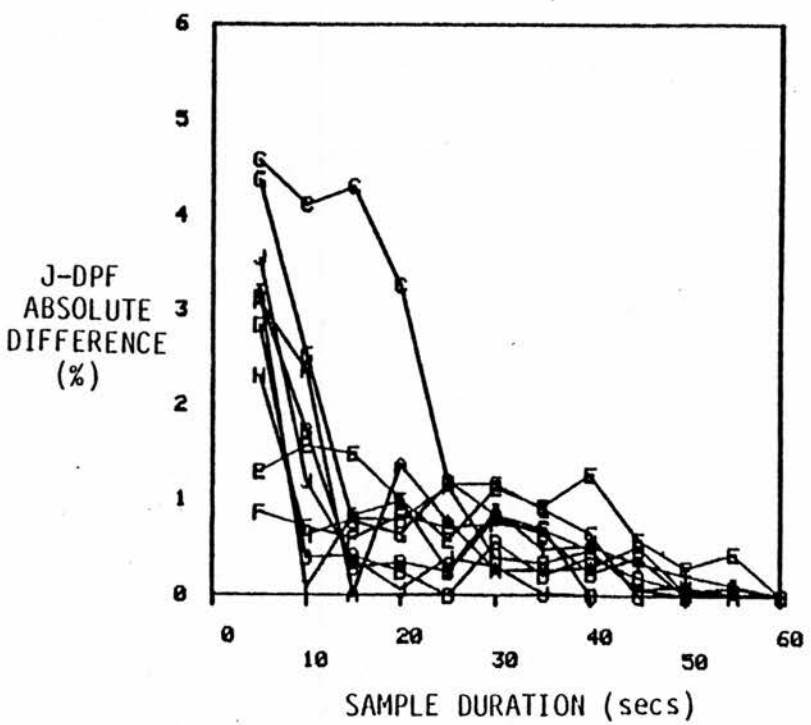


Figure 6.24 Absolute differences of J-DPF from the final long-term value plotted against sample duration (in cumulative 5 sec increments) for 10 healthy female speakers (labeled A-J).

final durations). As would be expected, the difference curves behave in a similar manner to the original curves in that the difference curves appear to stabilize around the 40 sec increment. One advantage of using the absolute difference curves is that all the parameters demonstrate decreasing values with increasing time thus permitting comparisons between the various parameters. For example, Figures 6.15 and 6.20 display the difference curves for the J-AVEX parameter in percent for the 2 groups of speakers (the vertical scales of these 2 figures also differ due to the larger difference values for some of the female speakers). As can be seen in Figs. 6.15 and 6.20, the J-AVEX parameter is now represented as differences in percent which decrease in value as the time increment approaches the final durational values for J-AVEX.

A further advantage found for the difference curves is that threshold tests can be applied to determine acceptable durational stability for each parameter. A threshold can be defined as some agreed proportion of the final long-term value of a given parameter. It is best expressed in this case as a percentage of the final value, rather than as a single absolute value of F0 in Hz, in order to normalize between different speakers. Another consideration is the proportion of members of the group of subjects whose difference curves successfully pass a given threshold. It was decided that 95% of the subject group would form a suitably representative proportion. Thus, the conjunction of the two criteria allows the minimum duration for group-stability to be established.

As a further condition, it was decided that having passed a given percentage threshold, each difference curve should remain within that band for at least 10 secs further before it could be considered adequately stable. This condition ensures some stability in a parameter's behavior for a given threshold. The problem of the differing durations of the various speakers' speech samples is partly addressed by this stability condition.

The threshold criteria are summarized as follows: 1) a difference curve is derived for each speaker on each parameter based on the differences between each cumulative value and the speaker's final long-term value, 2) a threshold is determined for each difference curve based on a percentage of the final long-term value for each curve, 3) a duration for each parameter is chosen based on the requirement that 95% of the speakers pass a given threshold and 4) the choice of threshold is conditioned by the requirement that the difference curve must remain within the given threshold band for at least 10 seconds. Figures 6.25 and 6.26 are examples of using the threshold criteria for the F0-AV parameter for the male and female groups, respectively. Each Figure is divided into 2 sections where section (a) displays the results for the application of the 1% of final value threshold and section (b) displays the results for the 2% of final value threshold. Each section of the Figures is in the form of a bar graph, the abscissa representing sample duration in cumulative 5 second increments and the ordinate representing each individual speaker (labeled A - J as in the previous Figures). Each bar depicts when the speaker passed a threshold (i.e. the transition from the unshaded to the shaded region) and the duration over which the difference curve remained within the threshold level

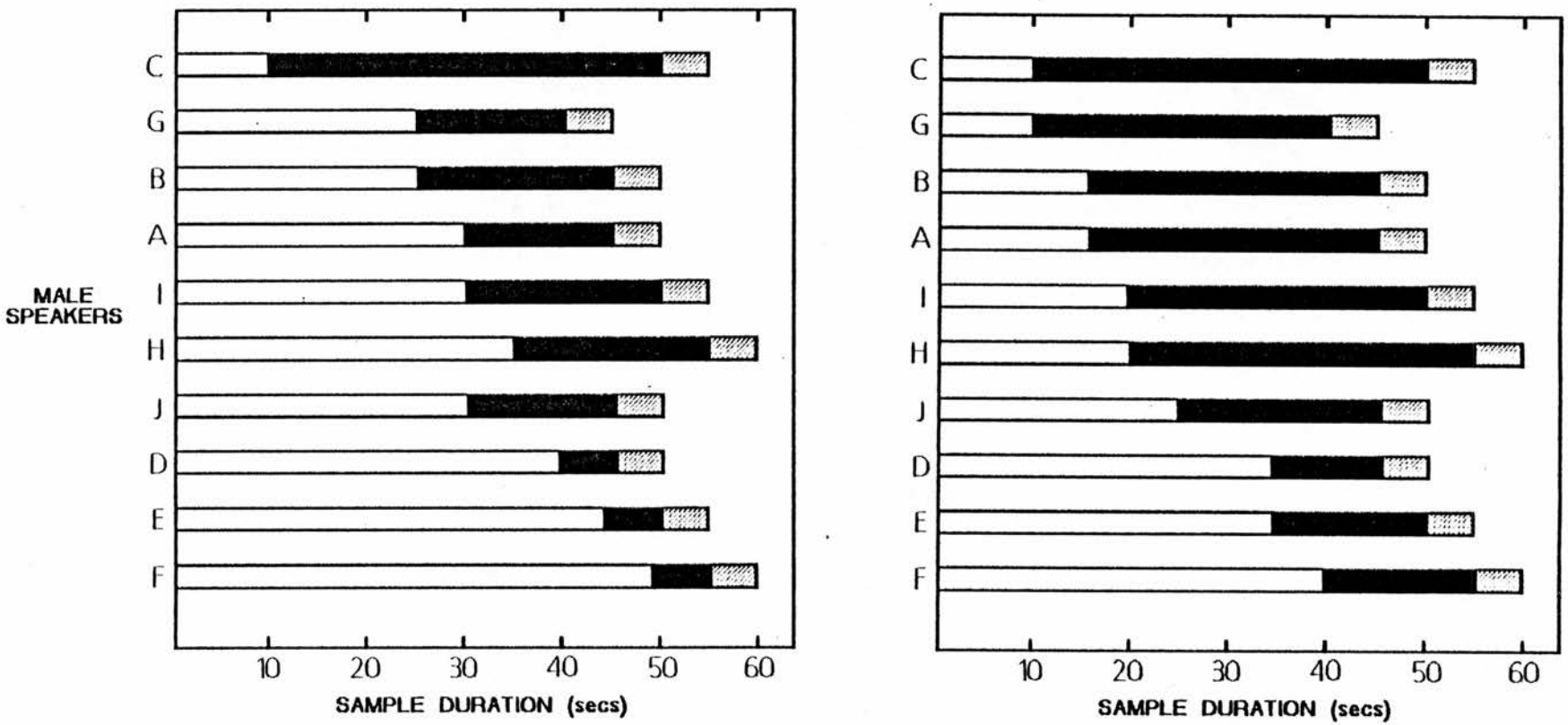


Figure 6.25 Bar graphs displaying the sample durations at which the male speakers pass 1% or 2% thresholds for FO-AV (Hz) and the periods during which speakers remain below the threshold. Each bar is divided into 3 sections: Unshaded -- sample durations before passing the threshold; shaded -- sample durations below the threshold; hatched -- marks the overall sample duration for each speaker. (a) -- 10 healthy male speakers (A-J) for the 1% threshold; (b) -- 10 healthy male speakers for the 2% threshold.

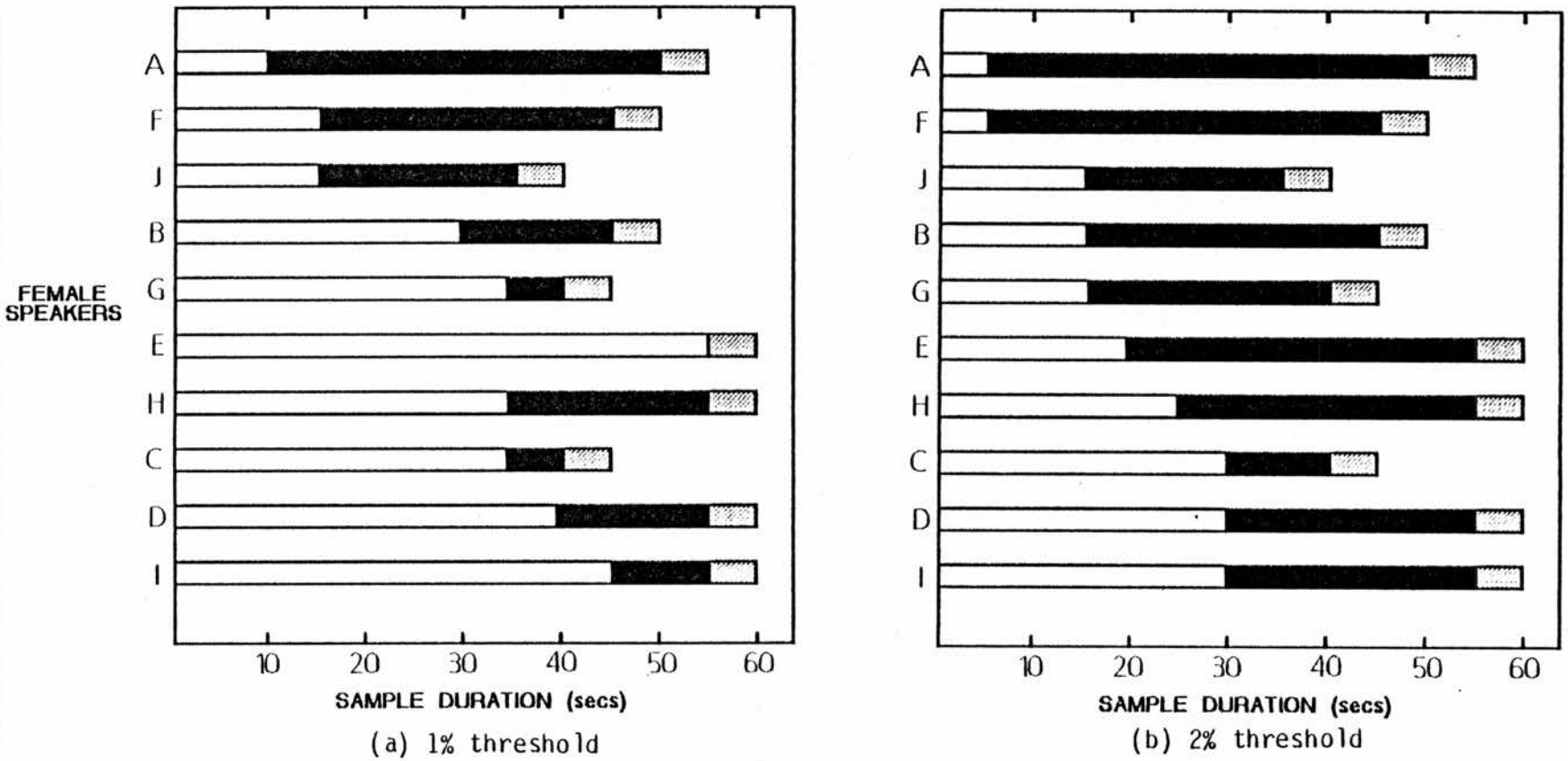


Figure 6.26 Bar graphs displaying the sample durations at which the female speakers pass 1% or 2% thresholds for FO-AV (Hz) and the periods during which speakers remain below the threshold. Each bar is divided into 3 sections: Unshaded -- sample durations before passing the threshold; shaded -- sample durations below the threshold; hatched -- marks the overall sample duration for each speaker. (a) -- 10 healthy female speakers (A-J) for the 1% threshold; (b) -- 10 healthy female speakers for the 2% threshold.

(i.e. the transition from the shaded region to the hatched region). The final point in the hatched region marks the overall duration of the speaker's speech sample. The differing thresholds displayed in the figures (i.e. 6.25a and 6.26a versus 6.25b and 6.26b) represent 2 attempts to determine when 95% of the speakers fell within a given percentage of their final F0-AV values. For example, male speaker H in Figure 6.25a displayed a F0-AV difference curve which passed a 1% of final value threshold at 35 secs and remained within that band for 15 seconds. For the 2% of final value threshold shown in Figure 6.25b, male speaker H passed that threshold at 20 secs and remained within that band for 35 seconds. If all the F0-AV difference curves in Figures 6.25a and 6.26a are examined in this way, one can see that only 70% of the male and female speakers fall within the 1% of final value threshold for the required 10 sec duration while all the speakers fulfill the requirements at the 2% of final value threshold (as shown in Figure 6.25b and 6.26b). Applying the 95% criterion for group performance to the 2% of final value bars of the male and female speakers, it appears that a minimum 35 sec speech sample would be sufficient to derive long-term F0-AV values to within a 2% of final value accuracy of each speaker's long-term behavior. Therefore, this threshold method suggests the appropriate duration for a given parameter for a given accuracy.

Table 6.1 summarizes the threshold findings for all the parameters examined in this study. The F0-AV and F0-DEV measures required a 2% of final value threshold to achieve at least 95% agreement amongst the 20 speakers. A duration of 35 secs of oral reading fulfilled all the requirements for the F0-AV for a combined group of 10 male and 10 female speakers. The table also includes a

Parameter	N	Threshold (%)	Percent Agreement	Seconds
FO-AV	10M + 10F	1	70	NA
		2	100	35
FO-AV	10M	1	70	NA
		2	100	35
FO-AV	10F	1	70	NA
		2	100	30
FO-DEV	10M + 10F	2	70	NA
		5	100	35
FO-DEV	10M	2	80	NA
		5	100	35
FO-DEV	10F	2	90	NA
		5	100	35
J-AVEX	10M + 10F	5	70	NA
		10	95	35
J-DEVEX	10M + 10F	5	85	NA
		10	100	30
J-RATEX	10M + 10F	5	95	40
		10	100	25
J-DPF	10M + 10F	5	85	NA
		10	100	25

Table 6.1 This table indicates the sample durations required for each parameter to reach stability based on the application of the threshold and group agreement criteria. NA -- not applicable since group did not fulfill 95% group agreement criterion.

breakdown of intonational results into 2 groups based on gender. The results for the two subgroups are similar though the female group required only 30 secs for stabilization of F0-AV as compared to the 35 sec duration of the male group. Three of the 4 perturbation measures (J-AVEX, J-DEVEX and J-DPF) reached the 95% group agreement for a threshold of 10% of final value while the J-RATEX measure obtained 95% group agreement at the 5% of final value threshold level. The lower threshold level reached by the intonational measures as compared to the perturbation measures reflects a more rapid approach and stabilization of the difference curves towards the final long-term intonational values. Therefore, more speech data would need to be elicited from a number of the speakers to produce long-term perturbational measures with the accuracy demonstrated for the intonational measures. Having noted this limitation, it can be seen that durations of 25 to 40 secs will produce perturbation measures with substantial accuracy for samples of oral reading from healthy speakers. Thus a duration of 40 seconds can be considered a useful practical value for evaluating perturbation and intonation measures derived from samples of oral reading produced by healthy speakers. It remains to be established by further research whether this duration would give equally satisfactory results in the evaluation of perturbation and intonation measures derived from speakers with laryngeal pathology. In addition, this study should be extended to the other parameters based on amplitude perturbations found in samples of connected speech. Based on the findings for the frequency perturbation parameters, it is suggested that 40 seconds of oral reading will suffice for the analysis of amplitude perturbations as well. To fully evaluate the effect of the increased sampling rate on the

analysis of speaker-characterizing perturbations in connected speech, this study should be replicated for the 10 male speakers but with their voice samples digitized at 20 KHz. Further research should examine the durational aspects of the perturbation measures for a variety of speaking tasks including differing speech samples and replicability of a given task.

SECTION 6.1.5 -- CONCLUSIONS OF THE DURATIONAL STUDY

Forty sec samples of read speech should provide relatively stable long-term speaker-characterizing parameters of intonation and perturbation in healthy speakers. This finding is in general agreement with the results of previous studies of the long-term features of the voice. Comparable durations of speech samples are required to produce stable long-term voice parameters from healthy female and male speakers. For the highest levels of accuracy, perturbation measures would require longer durations of speech than do intonational measures. Further research is required to determine these durations for the more accurate perturbation measures. The development of durational growth curves for the intonational and perturbational measures may provide a useful indicator of voice function.

SECTION 6.2 -- EFFECTS OF LOW-FREQUENCY PHASE COMPENSATION IN PERTURBATION ANALYSIS

One possible source of pitch period extraction error is the low-frequency phase distortion of a speech signal which has been recorded by an analog magnetic tape recorder. Distortion of the input speech waveform is caused by the reactive and resistive

components in the tape recording and playback system which do not maintain the relative phases of the harmonics of the recorded signal (Olsen 1982). Most of the phase distortion arises from recording and playback amplifiers and pre-amplifiers while the distortions from a good recording microphone are considered negligible (Berouti et al. 1977). Upon playback, the observed temporal structure of the signal has been altered though the overall periodicity of the waveform is approximately preserved (Berouti et al. 1977; Hess 1983).

A number of speech analysis techniques require speech data which has not been phase distorted, for example, the accurate determination of the closed phase of the glottal vibratory cycle and the derivation of the glottal volume velocity waveform via inverse filter techniques (Holmes 1975; Berouti et al. 1977; Hess 1983). In the case of pitch analysis, extractors should produce accurate parametric data since the various periodic components of voiced speech are preserved in the phase distorted waveform. This should be true for extractors operating in the spectral domain (e.g. cepstral analysis) where the amplitude spectrum of voiced speech is not unduly affected by phase distortion. There are a number of instances when a need for undistorted speech data may be indicated for pitch detectors operating in the time domain. Firstly, a strong case has been made for the standardization of speech material to be used in comparisons of the speech analysis results within and between laboratories (see, for example, Holmes 1975; Rabiner et al. 1976; Hess 1983). Standardization of recorded speech materials requires that distortion arising from differing recording systems be eliminated. Secondly, phase distortion may affect the performance

of pitch period detectors which examine the overall temporal structure of the waveform, in particular, those extractors which use waveform peak maxima/minima and zero crossings as basic markers of periodicity (some examples of these PDAs include Dolansky 1954; 1955; Anderson 1960; Reddy 1967; Gold and Rabiner 1969; Miller 1975; Tucker and Bates 1978). It has been frequently observed that the asymmetry of the waveform in voiced speech is largely due to the presence of the first harmonic and therefore low-frequency distortions are likely to influence the polarity of the principal peak in an unpredictable manner (Hess 1983). Further, many pitch period detectors examining the temporal structure of the waveform use exponential decay functions as the basic extractor of pitch period peak maxima and minima (e.g. the time domain PDAs created by Dolansky 1954; 1955; Anderson 1960; Gold and Rabiner 1969). Askenfelt and Hammarberg (1981) reported that variations in amplitude from cycle-to-cycle introduced "pseudo-frequency perturbations" into the pitch contour since the decay function could not always follow rapid changes in peak amplitude. These amplitude variations were treated as vocally-produced shimmer in the waveform, but additional phase distortion of the recorded waveform could exacerbate pitch period detection anomalies associated with changes in peak amplitudes. In addition, measures of shimmer are typically based on peak amplitudes in the waveform (see, for example, von Leden and Koike 1970; Davis 1976; 1979; Kitajima and Gould 1976; Koike et al. 1977; Deal and Emanuel 1978; Horii 1980; Horii 1982; Ramig and Ringel 1983) and should preferably be extracted from undistorted speech data.

Methods of limiting low-frequency phase distortion in recordings include the use of purpose-built hardware and phase compensation techniques. Phase distortions are eliminated from the recording process itself by the use of hardware such as analog FM tape recorders or digital Pulse Code Modulated tape recorders. However, this type of recording equipment is usually found in well-equipped recording studios rather than in environments such as hospitals and schools. In these cases, standard magnetic tape recording facilities are often available for collecting speech samples. Phase compensation techniques as suggested by Holmes (1975) and Berouti et al. (1977) are useful for limiting phase distortion in tape recordings. Two phase compensation techniques are described by Holmes (1975). The first technique involves rerecording the waveform, in reverse, through a recorder similar to the original recording device. This reverse recording should cancel the effects of the original recorded phase distortions. In the second technique, the overall total amount of phase delay (i.e. the distortion) is determined for the system and a special filter (analog or digital) is used to compensate for the delay. It is very useful to have recorded a calibration tone of known structure (e.g. low-frequency square or triangular waves) on the tape when different tape recorders are used for recording and playback. The calibration tone should reflect the amount of distortion from the system after recording and playback as well as the success of the phase compensation procedure. If phase compensation is completed digitally, then further quality degradation of the tape is avoided. Berouti et al. (1977) used frequency domain techniques to compensate the phases in recorded speech waveforms. An FFT transforms the waveform to its associated magnitude and phase

(a) Input wavefrom

(b) Distorted waveform

(c) Phase compensated waveform .

Figure 6.27 Example of phase compensation of a 70 Hz triangular waveform. (a) original input waveform; (b) distorted waveform following recording and playback on a REVOX A77 recorder; (c) phase compensated waveform (compensation factor approximately equal to pi/2).

Pre-amplifier	Input Recorder	Output Recorder	Compensation Factor (%)
Built-in	Revox A77	Revox A77	55
Built-in	Ferragraph Logic 7	Ferragraph Logic 7	51
AKB-11 Power Supply + Built-in	Revox A77	Revox A77	80
Shure Mixer	Revox A77	Revox A77	63
Balanced Pre-amplifier	Revox A77	Revox A77	40
Built-in	Uher	Revox A77	55 *
AKB-11 Power Supply + Built-in	Ferragraph Logic 7	Revox A77	50

Table 6.2 Recording set-ups and associated phase compensation factors.

* noticeable amplitude distortion following phase compensation due to non-linear response of input recorder.

spectra -- the phases in the spectrum are compensated by the addition of a compensation factor representing the total phase delay in the system. An inverse FFT transforms the phase compensated spectra back to the time domain. A calibration tone is also very useful for this phase compensation technique. Holmes (1975) and Berouti et al. (1977) found that the major compensation required for tape recordings was for phase distortion while the need for magnitude compensation was considered to be negligible.

An example of phase compensation of a 70 Hz triangular wave using the technique of Berouti et al. (1977) is shown in Figure 6.27. Figure 6.27a displays the original triangular wave and Fig. 6.27b shows the distorted waveform following recording/playback and digitization. A REVOX A77 tape recorder was used for this example for 6.27b. Waveforms 6.27a and b were low-passed filtered at 10 KHz to prevent aliasing of the signals and sampled at a rate of 20 KHz. It can be seen in Figure 6.27a that the filtering and sampling had very little effect on the shape of the original triangular wave. Figure 6.27c displays the phase compensated waveform which required a compensation factor of approximately pi/2, similar to the findings of Berouti et al. (1977) for the REVOX A77. This compensation factor was found to be the same for a number of REVOXs tested in the laboratory. No magnitude compensation was used for this waveform. Table 6.2 presents compensation factors for a number of recording/playback systems. Each compensation factor is presented as a percentage of the phase factor derived by Berouti et al. (1977). The phase factor, PF, is inversely proportional to the frequency (f) multiplied by a constant (k):

$$PF = k/f$$

where PF = pi

when f = 45 Hz

Therefore, the actual compensation factor for a given recording system will be the percentage of the phase factor required to compensate for a given waveform. It should be noted that some of the systems presented in Table 6.2 include an external pre-amplifier for the microphone. In these systems, the built-in microphone pre-amplifier has been bypassed unless specified otherwise.

As can be seen from the above discussion, the effects of low-frequency phase distortions of speech waveforms caused by tape recording equipment are of concern to the estimation of pitch data by pitch detection algorithms. In the present study, low-frequency phase distortion of speech data may affect the perturbation measurement system in two ways. Firstly, the voice samples to be input to the measurement system will be recorded on a variety of recording systems. It would useful to limit variations in perturbation measurement results due to differences in type of tape recorder thus achieving a reasonable degree of voice sample standardization for the present analysis purposes. Phase compensation procedures may be necessary in a voice analysis system in which samples recorded at a variety of clinics are sent to a central voice analysis center for evaluation of voice quality by acoustic processing. Secondly, the initial fundamental frequency data input to the perturbation measurement algorithms is provided by a parallel processing PDA operating in the time domain. The basic extractor of this PDA uses peak minima and maxima in the waveform as anchor points to determine pitch periods. Six functions of periodicity based on the peak minima and maxima are examined by an

exponential decay function and the results of this extraction procedure will eventually lead to the selection of a pitch period estimate. Therefore, pitch estimation by this PDA is in the first instance dependent on the amplitude structure of the time domain waveform and the addition of low-frequency phase distortion to the input speech signal may unduly affect the pitch extraction results. For the perturbation measurement system, both jitter and shimmer parameters may be affected by low-frequency phase distortion. It seems reasonable to determine if the compensation of recorder-induced low-frequency phase distortion affects the measurement of intonation and perturbation parameters from voice samples produced by healthy and pathological speakers. Of particular interest is the effect of phase compensation on the ability of the perturbation parameters to differentiate between healthy and pathological speakers. The purpose of the following experiment is to determine the effects of low-frequency phase compensation techniques on the analysis of intonation and perturbation parameters derived from the speech of a group of healthy speakers as well as a group of pathological speakers diagnosed as having a variety of epithelial disorders of the larynx.

SECTION 6.2.1 -- SPEAKERS, SPEECH MATERIAL, INSTRUMENTATION, PROCEDURES AND STATISTICS

Twenty adult male speakers were divided into two groups for this experiment. One group consisted of 10 speakers (aged 19 to 63 years, mean = 36.4 years) who reported no disorders of the larynx -- this group is referred to as the HEALTHY group. None of the speakers in the healthy group reported a history of speech or

phonatory disorders. Smokers were not excluded from this investigation. The other group contained 10 speakers (aged 32 to 82 years, mean = 60.6 years) who evidenced a variety of laryngeal disorders involving disruption of the epithelial tissues of the vocal folds. Table 6.3 lists the types of epithelial disorder and the number of speakers who evidenced each particular disorder in what is referred to as the PATHOLOGICAL group. All the pathological speakers were found to have a unilateral lesion of the epithelial tissues except for one case of papilloma in which the mass was located at the anterior commissure of the glottis. Epithelial disorder was chosen as a grouping factor since it was thought likely that substantial differences in fundamental frequency and perturbation parameters would be found between the healthy and pathological groups. The degree to which these two groups differed based on the use of phase compensation techniques should be a useful indicator of the need for these procedures. In addition, the detection and discrimination of epithelial disorders such as laryngeal carcinoma is a primary concern of the perturbation measurement system in general.

Each speaker completed an oral reading of the first two paragraphs of the "Rainbow Passage" (Fairbanks 1960). Prior to the reading, each speaker familiarized himself with the passage and was asked to read at a comfortable loudness level and rate.

The following equipment was used to tape record the voice samples produced by each speaker:

1) 10 speakers in the healthy group:
   Shotgun microphone (Sennheisser MKH815T) with power supply (Audio Engineering AKB11) --> REVOX A77 recorder

2) 7 pathological speakers:

NO. OF SPEAKERS	DISORDER
5	SQUAMOUS CARCINOMA (1 case of vocal fold fixation)
2	PAPILLOMA
1	VERRUCOUS TUMOR
1	HYPERKERATOSIS
1	ERYTHROLEUKOPLAKIA

Table 6.3 Types of epithelial disorder diagnosed for the 10 speakers in the pathological group. All speakers evidenced a unilateral placement of the epithelial disorder except for 1 case of papilloma where the disorder was located at the anterior commissure of the vocal folds.

Dynamic Cardioid Microphone (Sennheisser MD421) —> Microphone Mixer/Amplifier (Shure M67-2E) —> REVOX A77 recorder

3) 3 pathological speakers:
Dynamic Cardiod Microphone (Sennheisser MD421) —> BALANCED pre-amplifier --> REVOX A77 recorder

Each recording was completed in a sound-treated booth to limit background noise. A REVOX A77 tape recorder was used to play back the voice samples for digitization on to a computer.

Each voice sample was processed for fundamental frequency and perturbation data as described in Section 5.5 above. Two analysis conditions were applied to each voice sample including 1) phase compensation of the signal prior to the perturbation analysis and 2) no compensation of the speech waveform. The phase compensation technique used in this experiment was the frequency domain procedures of Berouti et al. (1977) as described in the introduction to this section. For each voice sample, only the phase spectrum was compensated while the magnitude spectrum was not manipulated. Table 6.2 contains the phase compensation factors used for the 3 recording/playback systems described in the instrumentation section of this experiment. It should be noted that these factors compensate for phase distortion which might arise from the anti-aliasing filter and the analog-to-digital convertor as well as from the tape recorders. The minimal low-frequency phase distortion effects due to the recording microphones and the digital low-pass filter used in the PDA pre-processor are not accounted for by the compensation factors.

The following 10 fundamental frequency and perturbation parameters were evaluated by statistical analyses: F0-AV, F0-DEV, J-AVEX, J-DEVEX, J-RATEX, J-DPF, S-AVEX, S-DEVEX, S-RATEX AND S-DPF. Group means and standard deviations were determined for the 10 parameters for the following four group comparisons:

A. Within-Group Comparisons
   1. Healthy Group/No Compensation vs. Healthy Group/Phase Compensation
   2. Pathological Group/No compensation vs. Pathological Group/Phase Compensation

B. Between-Group Comparisons
   1. Healthy Group/No Compensation vs. Pathological Group/No Compensation
   2. Healthy Group/Phase Compensation vs. Pathological Group/Phase Compensation

Comparisons A1 and A2 are included in the study to determine whether or not phase compensation techniques in general significantly affect the measurement of F0 intonation and perturbation parameters for either group of speakers. In comparisons B1 and B2, two effects are under investigation. Firstly, for either comparison B1 or B2, it is expected that a number of the parameters will be significantly different between the two groups of speakers. In particular, it is expected that the group evidencing the epithelial disorders will demonstrate greater than normal levels of irregular laryngeal vibration as revealed by increased perturbation measures due to the asymmetrical location of growths in the vocal folds of these speakers. Secondly, if phase compensation does affect the measurement of fundamental frequency and perturbation parameters then this may alter the separation of the two groups as shown in the results for comparisons B1 and B2. Each of the ten parameters were investigated for significant group differences for the 4 conditions by a Student's t test for comparing the two mean values of each

group.

SECTION 6.2.2 -- RESULTS AND DISCUSSION

The group means and standard deviations are presented in Table 6.4 for each of the fundamental frequency and perturbation parameters. Each column of the table contains the distributional results for each of the 4 experimental conditions including 1) healthy group with no compensation (HO), 2) healthy group with phase compensation (HP), 3) pathological group with no compensation (PO) and 4) pathological group with phase compensation (PP). Table 6.5 displays the results of statistical tests of significant differences between the four conditions based on Student's t tests for the 10 parameters. The within-group comparisons include HO vs HP and PO vs PP while the between-group comparisons include HO vs PO and HP vs PP. For any given parameter within any given comparison of groups, a significant difference was reached at the a = .05 level. In Table 6.5, a significant difference is marked with reference to the first group in a given comparison by the less-than marker (the first group is significantly < the second group for a given parameter) or the greater-than marker (the first group is significantly > the second group). Non-significant differences between groups for a given parameter are denoted by NS.

A. Within-Group Results (HO vs HP; PO vs PP): Comparisons of the HO column to the HP column of Table 6.4 highlight some trends for the healthy group in regard to the effects of phase compensation techniques on the measurement of FO and perturbation parameters. Firstly, it would appear that phase compensation has very little

PARAM.	HO MEAN	HO SD	HP MEAN	HP SD	PO MEAN	PO SD	PP MEAN	PP SD
FO-AV	113.90	14.69	113.64	14.41	139.31	23.53	139.09	24.37
FO-DEV	23.19	5.79	23.26	5.84	29.72	10.75	29.73	10.34
J-AVEX	15.90	2.36	15.84	2.63	17.32	5.07	17.40	5.11
J-DEVEX	5.30	0.51	5.22	0.79	7.22	3.01	7.31	3.07
J-RATEX	25.59	2.29	23.22	2.25	33.75	11.73	33.42	12.61
J-DPF	19.00	3.35	15.98	2.63	25.73	8.01	24.82	8.49
S-AVEX	59.78	36.12	56.92	18.16	215.17	425.66	166.50	158.18
S-DEVEX	18.57	2.97	16.53	2.68	23.91	12.65	21.40	6.42
S-RATEX	61.92	4.62	58.14	5.45	64.04	12.20	63.16	12.23
S-DPF	29.05	5.04	25.21	5.00	35.08	8.92	34.60	8.77

Table 6.4 Group means and standard deviations (SD) for the 10 intonation and perturbation parameters for the four conditions:

HO -- healthy group with no compensation
HP -- healthy group with phase compensation
PO -- pathological group with no compensation
PP -- pathological group with phase compensation.

	WITHIN GROUP		BETWEEN GROUP	
PARAM.	HO:HP	PO:PP	HO:PO	HP:PP
FO-AV	NS	NS	.01 <	.02 <
FO-DEV	NS	NS	NS	NS
J-AVEX	NS	NS	NS	NS
J-DEVEX	NS	NS	NS	NS
J-RATEX	.05 >	NS	.05 <	.05 <
J-DPF	.05 >	NS	.05 <	.01 <
S-AVEX	NS	NS	NS	.05 <
S-DEVEX	NS	NS	NS	.05 <
S-RATEX	NS	NS	NS	NS
S-DPF	NS	NS	NS	.01 <

Table 6.5 Results of Student's T tests for 10 intonation and perturbation parameters for four comparisons of group behavior. NS — nonsignificant difference; > -- the first group value is significantly greater; < -- the first group is significantly less.

HO — normal group with no compensation
HP -- normal group with phase compensation
PO — pathological group with no compensation
PP -- pathological group with phase compensation

effect on the long-term fundamental frequency parameters FO-AV and FO-DEV for the healthy group. Secondly, six of the perturbation parameters (J-RATEX, J-DPF, S-AVEX, S-DEVEX, S-RATEX AND S-DPF) display a trend of slightly lowered group mean values for the phase compensated condition as compared to the no compensation condition. The HO vs HP column of Table 6.5 presents the results of statistical tests of significance between the 10 parameters for the two compensation conditions of the healthy group. It can be seen that for most of the parameters there are no significant differences for the HO versus the HP condition. However, two perturbation parameters, J-RATEX and J-DPF, demonstrate significantly greater group values for the no compensation condition as compared to the phase compensation condition. The use of phase compensation on the acoustic recordings of the healthy group voice samples resulted in lower measures of perturbation. It is suggested that the significantly lower jitter scores reflect the general trend for lower shimmer parametric values in the phase compensation condition -- the use of phase compensation to decrease the perturbation of the acoustic waveform amplitude structure due to the tape recording/playback system is evidenced as decreased jitter measures since the parallel processing PDA relies on amplitude measures to make pitch period estimates.

The columns labeled PO and PP in Table 6.4 present the group means and standard deviations for the 10 acoustic parameters derived from the voice samples of the pathological speakers. As in the case of the healthy group, it would appear that phase compensation has very little effect on the group averages for the long-term measures FO-AV and FO-DEV. There is also a trend towards lower group mean

values for a number of the perturbation parameters (J-DPF, S-AVEX, S-DEVEX, S-RATEX AND S-DPF) for the phase compensated condition as compared to the no compensation condition of the pathological group (although some of these differences are rather small). Table 6.5 presents the results for within-group tests of significance for the pathological group in the column labeled PO vs PP. It can be seen that no significant differences were found for any of the parameters when phase compensation was used for the pathological voice samples. The lack of significant differences in the pathological group is caused by the spread of values for each parameter (i.e. the standard deviation scores in columns PO and PP of Table 6.4) as compared to the healthy group spreads. That is, the pathological group is not as homogeneous in laryngeal behavior as the healthy group and therefore the effects of phase compensation may be masked for this group of dysphonic speakers. In addition, for any given dysphonic speaker, the increased waveform perturbations associated with a given epithelial disorder may be random enough in its own right such that fine adjustment of the speech waveform by phase compensation does not significantly alter the perturbation results. It should be noted, however, that at least one parameter was substantially affected by the addition of phase compensation for the pathological group -- S-DEVEX demonstrated a large decrease in value for the pathological speakers as seen in Table 6.4.

B. Between Group Results (HO vs PO; HP vs PP): The group mean and standard deviations for the 10 parameters presented in Table 6.4 can also be used to compare the two groups of speakers. In the no compensation conditions (HO vs PO), there is a trend towards higher FO-AV and greater FO-DEV for the pathological group as compared to

the healthy group fundamental frequency measures. In addition, all 8 perturbation parameters demonstrated higher group means for the pathological speakers relative to the results of the healthy group. It can be seen in Table 6.5 that significant differences were found between the two groups for 3 parameters in the no compensation condition -- F0-AV, J-RATEX AND J-DPF parameters are significantly higher for the pathological group as compared to the healthy group. As a group, the dysphonic speakers with epithelial disorders of the vocal folds demonstrated higher long-term average fundamental frequency and increased waveform jitter relative to the phonatory behavior of speakers with normal laryngeal structures. The increased average fundamental frequency values may be related to increased stiffness of the vocal fold epithelial tissue associated with the various disorders. The location of the epithelial disorders was unilateral in 9 out of 10 of the dysphonic speakers and therefore produced asymmetrical vibratory patterns of the vocal folds and the consequent increased jitter measures (Mackenzie et al. 1983). It is encouraging that the two groups of speakers can be differentiated by the intonation and perturbation parameters though only a few of the measures demonstrated significant differences.

The effect of phase compensation of the acoustic voice samples can be observed for the between-group results as displayed in Tables 6.4 and 6.5. The trends for increased values of fundamental frequency and perturbation parameters for the pathological group as compared to the healthy group are also found in the phase compensation conditions. The following observations have been noted for the HP and PP conditions as presented in Table 6.4. Firstly, phase compensation of the voice samples does not appear to have

greatly affected the differences in fundamental frequency parameters F0-AV and F0-DEV between the two groups of speakers when compared to the no compensation between-group differences. Three perturbation parameters (J-AVEX, J-DEVEX and S-DEVEX) demonstrated very little change in between-group differences with the addition of compensation. Secondly, 4 of the perturbation parameters (J-RATEX, J-DPF, S-RATEX AND S-DPF) demonstrated increased differences in mean values between the two groups in the phase compensation condition as compared to the no compensation condition. Finally, the shimmer measure S-AVEX revealed a reduced difference in between-group values in the phase compensation condition.

Table 6.5 presents the results of the tests of significance for the 10 parameters for the between-group phase compensation comparison (HP vs PP). Here, it can be seen that 6 of the parameters demonstrated significantly different results between the two groups of speakers. The fundamental frequency parameter F0-AVEX is significantly higher for the pathological group as compared to the healthy group. The following perturbation measures were found to be significantly greater for the pathological group in relation to the healthy group: J-RATEX, J-DPF, S-AVEX, SDEVEX AND S-DPF (the strength of the difference for two of these parameters, J-DPF and S-DPF, is quite high). It is interesting to note that despite the decreased difference between the group values with the addition of phase compensation, the parameter S-AVEX revealed a significant difference between the two groups. These results for the phase compensation condition provide further support for the description of speakers with epithelial disorders of the vocal folds as being high in average fundamental frequency and waveform perturbations

compared to healthy speakers. The effects of phase compensation of the voice samples is quite clear since 6 out of the 10 parameters are significantly different between the two groups of speakers as compared to 3 significant differences in the no compensation between-group comparisons. It is reasonable to suggest that phase compensation techniques would be useful for the analysis of perturbations in samples of speech recorded from speakers with other types of phonatory disorder, particularly when an asymmetrical disruption of the vocal folds exists (e.g. unilateral vocal nodules and polyps).

The results of this experiment have some bearing on the practical application of the perturbation analysis system as a clinical tool for the evaluation of the voice. The phase compensation procedure is by far the most computationally expensive process in the perturbation measurement system as used in the present study. The phase compensation technique completes 2 Fourier analyses per interval of speech (approximately 410 ms of data) for a total voice sample of 40 secs. In addition, the phase compensated speech signal must then be low-pass filtered by a digital filter as the first step in the pitch detection process. Both of these procedures are time-consuming, particularly in comparison to simple low-pass pre-processing by a hardware filter prior to the analysis of perturbations. It is suggested that the perturbation measurement system may be used in one of two schemes (Hiller et al. 1983). Firstly, the voice evaluation system could be a centralized one, in which tape recorded voice samples are sent to a single facility for evaluation. In this case, the need for special hardware for preventing low-frequency phase distortion as suggested by Holmes

(1975) is highly recommended. This special device need only be used at the central processing facility and therefore more readily available tape recorders (of known specifications) could still be used in the field for collecting voice samples. In the second case, a microcomputer-based system could be used for local clinical evaluation of voice samples. Direct voice input to an analog-to-digital convertor connected to the microcomputer would prevent most low-frequency phase distortions and the consequent need for compensation techniques (the study of Zyski et al. 1984 is one example of this type of data acquisition system).

Finally, do the differing and often conflicting results found for perturbation studies in the literature reflect, in part, the differing effects of low-frequency distortion associated with the recording of voice samples? The results of the present study suggest the need for standardization of taped materials in this crucial area of voice analysis for the screening and differentiation of voice pathologies.

SECTION 6.2.3 -- CONCLUSIONS OF THE PHASE COMPENSATION STUDY

The results of the present experiment suggest the need to phase compensate for low-frequency distortions of voice samples associated with the tape recording/playback process prior to analysis for waveform perturbations. In the first instance, phase compensation will improve the performance of the parallel processor for pitch detection in the time domain since it uses amplitude measures derived from the temporal structure of the waveform which is affected by phase distortion. The results of the phase compensation

experiment demonstrated the importance of these techniques for the differentiation of speakers with epithelial vocal disorders from healthy speakers. The elimination of phase distortion by any means should lead to more standardized analysis and evaluation of perturbations derived from the speech of healthy and pathological speakers.

SECTION 6.3 — SPEAKER GROUP SELECTION AND EVALUATION

In this section, the discussion focuses on the characteristics of the speaker groups to be used in a number of parametric statistical analyses of the F0 intonation and perturbation measures. The discussion begins with the selection and grouping of acoustic data derived from the recordings of healthy and pathological speakers. The remainder of this section provides a description and analysis of the parameter sample distributions resulting from the speaker selection procedures. In the first instance, recorded samples of connected speech were available from a total of 238 speakers. Table 6.6 displays a breakdown of this initial data pool into six speaker groups where 2 of the groups, CA and PA, represent the condition of the larynx regardless of the sex of the speaker while groups CM, CF, PM and PF also reflect the sex of the speaker. The control speakers consisted mostly of students and staff members of the university. Each control speaker reported a healthy state of the larynx at the time of recording and no history of chronic throat illness (smokers were however not excluded from the data pool). The pathological speakers displayed a variety of disorders of the vocal folds as diagnosed by laryngologists. These speakers represent the typical outpatient populations as examined in two hospital voice

clinics. The recording elicited from each speaker was analyzed by the perturbation measurement system as described in Section 5.5 above and the results stored for statistical analysis. The following 10 parameters were analyzed for each of the six speaker groups: F0-AV, F0-DEV, J-AVEX, J-DEVEX, J-RATEX, J-DPF, S-AVEX, S-DEVEX, S-RATEX and S-DPF.

Initial observations of the speaker group behavior were based on distributions derived from each group on each of the 10 parameters. These parameter distributions displayed a number of characteristics. Firstly, unimodal distributions of the values were noted for each parameter in each speaker group. Secondly, for 9 out of the 10 acoustic parameters, each distribution was somewhat normal in appearance with a clearly discernible peak close to the mean value. However, the shimmer measure S-DEVEX displayed highly skewed distributions in the positive direction for each of the 6 speaker groups. Lastly, though the group distributions for the other 9 acoustic parameters were normal in appearance, a small number of individual data points were clearly separated from their associated distributions.

Based on these preliminary observations of the speaker groups, further analyses of the group distributions were completed for the selection of speakers to be used in other statistical procedures. Firstly, the shimmer parameter S-DEVEX was not used in the speaker selection procedure due to the highly skewed distributions evidenced for each of the 6 original speaker groups. Secondly, speakers were eliminated from their respective groups if they demonstrated large deviations in parametric values from the group means of their respective groups. That is, each speaker group was manipulated to

exclude those speakers with parameter values which were considered unrepresentative of the group as a whole. This manipulation of the data is, to a certain extent, a violation of the notion of random sampling (which may not exist in the first instance for a self-selected population as found in an outpatient voice clinic) but should ensure some normality in the sample distributions to be used in the parametric statistical analyses. Based on the means and standard deviations derived from the remaining 9 parameters, a speaker was eliminated from his respective speaker group if:

a) Any one or more of the parametric values were greater than 3 SDs from the mean values for the given group or

b) Any 3 or more of the parametric values were greater than 2.5 SDs from the mean values for the given group.

In the case of the groups CA and PA, the F0-AV and F0-DEV parameters were not used in the speaker selection procedure since it is expected that these measures are correlated with the sex of the speaker. The results of the selection procedures are presented in Table 6.7. As can be seen in this table, the elimination of individual speakers from their respective speaker groups varied depending on whether sex of the speaker was considered as a grouping factor.

The distributional data was then recomputed for each acoustic parameter using the 6 newly formed speaker groups. Table 6.8 presents the group means and SDs for each of the parameters derived for the 6 speaker groups -- note that the groups CA and PA do not have group F0-AV and F0-DEV values associated with them. Of primary interest is the general nature of the sample distributions of each parameter derived for each group following the selection procedures. In particular, do the sample distributions for each parameter appear

1. All Healthy Control Speakers (CA) - 129
    a. Control Males (CM) ---------- 70
    b. Control Females (CF) ---------- 59
2. All Pathological Speakers (PA) ---- 109
    a. Pathological Males (PM) ------- 55
    b. Pathological Females (PF) ----- 54

Table 6.6 Breakdown of initial data pools available for statistical tests.

1. CA ----- 124 (4 males and 1 female eliminated)
    a. CM - 66 (3 of the 4 males as in CA)
    b. CF - 58 (the same female as in CA)
2. PA ----- 106 (2 males and 1 female eliminated)
    a. PM - 51 (1 of the 2 males as in PA)
    b. PF - 52 (the same female as in PA)

Table 6.7 Final breakdown of speakers following selection procedures.

a) ALL CONTROL SPEAKERS (N=124)

PARAM	MEAN	SD	NORMAL χ^2_u	NORMAL χ^2_f	LOG χ^2_u	LOG χ^2_f	SKEW
JDEVEX	15.35	2.60	3.60(7)	5.89	4.24(7)	6.88	0
JAVEX	4.91	1.07	3.04(6	4.38	3.33(6)	4.86	0
SAVEX	14.91	3.47	5.23	9.69	5.04(6)	5.51	P
JRATEX	21.50	4.00	11.30(7)	19.37	12.44(7)	17.66	0
SRATEX	52.32	7.97	6.55(8)	7.28	8.81	12.95	N
JDPF	14.52	3.31	3.28(7)	3.30	2.43(7)	3.60	P
SDPF	24.01	4.78					

b) CONTROL MALE SPEAKERS (N=66)

PARAM	MEAN	SD	NORMAL χ^2_u	NORMAL χ^2_f	LOG χ^2_u	LOG χ^2_f	SKEW
FOAV	112.37	12.16	4.61	7.12	3.51	5.94	P
FODEV	20.49	5.91	5.05	9.96	2.17	6.55	P
JDEVEX	15.75	2.92	8.91(6)	11.32	9.48	14.87	N
JAVEX	5.01	1.11	.40(6)	2.03	2.08(6)	4.34	0
SAVEX	16.02	3.51	5.57	8.55	3.19	4.53	P
JRATEX	22.94	4.16	7.38(6)	14.81	5.18(6)	11.60	0
SRATEX	57.89	5.40	1.08(6)	1.93	1.88(6)	3.10	N
JDPF	16.21	3.26	6.96	9.00	3.99(4)	7.72	P
SDPF	26.67	4.45	4.24(6)	6.13	3.49	6.32	0

c) CONTROL FEMALE SPEAKERS (N=58)

PARAM	MEAN	SD	NORMAL χ^2_u	NORMAL χ^2_f	LOG χ^2_u	LOG χ^2_f	SKEW
FOAV	199.09	18.94	5.20	9.31	5.40	10.27	0
FODEV	39.54	6.03	4.74	10.14	3.47	8.72	P
JDEVEX	14.89	2.12	8.79	11.11	4.54(4)	14.52	P
JAVEX	4.81	1.03	2.27	3.53	1.36	3.13	P
SAVEX	13.72	3.10	3.34	5.03	3.15	5.14	0
JRATEX	20.05	3.67	.87(4)	3.91	1.42(4)	3.92	0
SRATEX	46.21	5.86	9.11	12.47	9.58	14.88	0
JDPF	12.80	2.68	10.41	12.92	10.47	12.88	0
SDPF	21.26	3.78	1.96	4.84	2.12	7.61	0

Table 6.8 Group statistics for each of the parameters derived for the 6 speaker groups. χ^2u -- χ^2 unforced; χ^2f -- χ^2 forced; P -- positive skew; N -- negative skew; 0 -- no skew. Continued on next page.

d) ALL PATHOLOGICAL SPEAKERS (N=106)

PARAM	MEAN	SD	NORMAL χ^2_u	NORMAL χ^2_f	LOG χ^2_u	LOG χ^2_f	SKEW
JDEVEX	16.12	4.61	7.55(6)	8.48	6.57(6)	9.14	P
JAVEX	5.78	2.45	11.04(7)	15.97	4.68(6)	10.71	P
SAVEX	19.04	7.17	8.58(7)	12.32	5.26	10.05	P
JRATEX	25.95	10.38	14.20(7)	17.85	11.91(6)	17.40	P
SRATEX	63.70	12.64	28.63(6)	38.88	43.40(6)	223.60	N
JDPF	19.37	7.69	10.92(7)*	13.06*	9.69(6)	18.44	P
SDPF	36.12	8.89	6.58(6)	18.20	12.43(6)	165.90	N

e) PATHOLOGICAL MALE SPEAKERS (N=51)

PARAM	MEAN	SD	NORMAL χ^2_u	NORMAL χ^2_f	LOG χ^2_u	LOG χ^2_f	SKEW
FOAV	119.98	19.64	10.59	11.31	7.56	7.90	P
FODEV	23.79	7.74	8.76	27.41*	2.68	9.78	P
JDEVEX	17.54	5.07	4.97	9.66	4.91	13.79	O
JAVEX	6.41	2.50	8.63	10.84	4.07	10.37	P
SAVEX	21.57	7.40	10.04	11.53	2.82(4)	7.44	P
JRATEX	29.58	9.75	4.45(4)	6.56	2.79	6.29	P
SRATEX	70.42	7.41	3.57	5.69	5.27	11.51	N
JDPF	22.97	6.83	3.33	7.84	1.35	10.77	P
SDPF	41.02	6.48	8.17	11.42	7.23	11.70	O

f) PATHOLOGICAL FEMALE SPEAKERS (N=52)

PARAM	MEAN	SD	NORMAL χ^2_u	NORMAL χ^2_f	LOG χ^2_u	LOG χ^2_f	SKEW
FOAV	185.47	27.27	5.75	8.46	3.53	8.24	P
FODEV	39.67	8.82	7.96	10.63	5.11(4)	9.55	P
JDEVEX	15.09	3.99	4.53	9.81	3.66(4)	10.96	O
JAVEX	5.29	2.23	5.61	6.26	1.25(4)	2.71	P
SAVEX	17.38	7.06	6.80(4)	13.32	6.54(4)	9.86	P
JRATEX	22.91	9.67	2.89(4)	10.44	2.83(4)	8.72	P
SRATEX	59.06	13.50	6.61	14.88	9.53	36.61	N
JDPF	16.36	6.89	5.52(4)	7.69	2.85(4)	5.87	P
SDPF	32.38	8.28	3.27	7.29	6.79	33.11	N

Table 6.8 Continued from previous page.

to be derived from populations with normal distributions? To answer this question, the chi-square statistic was applied to each sample distribution to determine the "goodness of fit" to a gaussian distributional curve. It should be noted that a chi-square test based on a gaussian distribution primarily evaluates the central intervals about the mean of the distribution since the expected frequencies are low in the tails of the gaussian curve (i.e. less than 5) and therefore not included in the calculation of the χ^2 value. The result of this restriction is the determination of χ^2 values in which the number of degrees of freedom (df) are relatively low compared to the actual number of intervals used to form a given sample distribution. For the present study, ten intervals were used to create the sample distributions which typically resulted in 5 df for the associated χ^2 estimate based on a gaussian curve fit. Based on the limited number of df, the following observations were made for each sample distribution:

a) If the χ^2 for a given distribution is significant then the central portion of the distribution is not representative of a normal population distribution -- the data points comprising this given sample distribution may require a transformation (e.g. by logarithm) to achieve a normal distribution or be used with caution in any further parametric statistical analyses.

b) If the resultant χ^2 for a given distribution is non-significant then it is assumed that the central portions of the data are representative of a normal population distribution.

Table 6.8 presents the results of the chi-square analyses of each parameter for each group based on a gaussian curve fit and limited number of df (labeled as χ^2u -- the unforced χ^2). For most of the sample distributions, the df is equal to 5 but differing values are listed in brackets next to the χ^2u value. A χ^2 value was

considered significant at a level $p<.01$ -- a star (*) denotes the significant results in Table 6.8. Only one sample distribution, the J-DPF parameter of the PA group, was found to be significant for this initial chi-square evaluation. Therefore, it is concluded that the majority of sample distributions derived for both pathological and control groups have central distributional values which are representative of normal statistical populations.

The results of the initial unforced chi-square tests are encouraging. However, due to the limited sample size of each group, χ^2 values based on relatively low numbers of df may not reflect possible irregularities in the tails of the sample distributions. To determine if such irregularities exist in the data, chi-square analyses were completed in which the tails of the gaussian distributions were forced into the computation of each χ^2 -- the result is a χ^2 value based on expected frequencies of less than 5, that is, 9 df for a 10 interval distribution. It is recognized that this manipulation of the chi-square statistical procedure produces a less powerful test of significance, but any gross deviations in distributional fit in the tails should be highlighted. As in the case of the unforced chi-square analyses discussed above, the following observations were completed for each sample distribution:

a) If a significant χ^2 is found then irregularities for each sample distribution exist in the tails of the distribution — the data may require an appropriate transformation or be used with caution in any further statistical analysis.

b) If the χ^2 is not significant then further evidence has been found for a normal population distribution as the source of the sample distribution.

The results of the forced chi-square tests (labeled as χ^2_f) for each sample distribution of each group is displayed in Table 6.8. In only 2 instances did a given sample distribution produce a significant forced χ^2 value -- the FO-DEV parameter of the PM group as well as the J-DPF of the PA group. Therefore, it is concluded that the majority of the sample distributions do not evidence irregularities in the tail regions and are representative of normal population distributions (this result is not surprising due to the original speaker selection procedures). The overall finding is that most of the sample distributions for the 6 speaker groups appear to be derived from statistically normal population distributions.

It was noted that many of the sample distributions were not perfectly symmetrical about their mean values. In Table 6.8, a one-word qualitative description has been given for each sample distribution in terms of skew -- positive (P), negative (N) and no skew (0). These descriptive statements of skew were based on visual observations of the sample distributions rather than from statistical evaluations. As can be seen from the table, a number of sample distributions have been described as positive in skew -- therefore a log normal curve fit was used in chi-square testing to determine if a logarithmic transformation of the sample values would be useful. The results of both unforced and forced chi-square tests based on a log normal distribution are also presented in Table 6.8 for each sample distribution in each speaker group. Note that χ^2 values are not presented for the negatively skewed data in which it is expected that a log transformation of the data would degrade any fit to a normal distribution. By comparing the results of the gaussian chi-square analyses to the log normal χ^2s, it can be

generally stated that log transformation of positively skewed sample distributions would produce slightly better fits to normal population distributions. One notable case is the F0-DEV sample distribution of the PM group which has non-significant χ^2 values when a log normal distribution is used. A number of the sample distributions labeled as not skewed also demonstrate some improved fits to the log normal distribution which would suggest some positive skews in the sample distributions.

The following is a summary of the speaker group analyses completed up to this point. Firstly, a number of speakers were eliminated from the original data pool due to extreme values on one or more of the acoustic parameters. The remaining speakers were pooled into 6 speaker groups, each group being evaluated for normality of sample distributions for each of the acoustic parameters. In a majority of cases, it can be assumed that the sample distributions are derived from normal population distributions. In addition, the transformation of a number of parameters by logarithmic techniques may be useful in further statistical analyses.

Some brief comparative comments are presented here for the distributional values found for each group of speakers. Firstly, the means and SDs of the two large groups, CA and PA, for the 7 perturbation parameters may be compared. Two observations of note are 1) the mean perturbation scores are greater for the pathological groups as compared to the control groups and 2) the standard deviations of the sample distributions are greater for pathological groups as compared to the control groups. Secondly, comparisons of the groups in which sex of the speaker has been used as a grouping

factor may also be made. Comparing across voice condition (i.e. CM vs PM and CF vs PF), one finds higher means and wider SDs for the pathological speakers as compared to the control speakers. However, the F0-AV and F0-DEV distributional values of the PF group are slightly lower as compared to the values of the CF group. Looking across the condition of sex within a given voice condition, it can be seen that, in general, the perturbation values of the male groups are greater than the female groups. Further discussion of these differences will be presented in the next section in which analyses of variance are completed for the data. Finally, Table 6.9 presents the disorders represented by the speakers in the pathological groups.

SECTION 6.4 — ANALYSIS OF VARIANCE OF THE FACTORS VOICE CONDITION AND SEX OF THE SPEAKERS

In Section 6.3, a number of speaker groups were defined and evaluated for the general distributional properties for a set of acoustic parameters. Two general trends were observed for the various sample distributions. Firstly, for the factor of voice condition, it was observed that higher group mean perturbation values are associated with the pathological groups (PA, PM and PF) as compared to their equivalent control groups of speakers (CA, CM and CF). Secondly, for the factor sex of the speaker, there was a trend for group mean perturbation values of the male speaker groups (CM and PM) to be higher than the averages displayed by their associated groups of female speakers (CF and PF). The aim of the present section is to further evaluate these two factors of voice condition and sex of speaker to determine if the discrimination

TYPE OF PATHOLOGY	MALES	FEMALES	ALL
DISORDERS OF THE LIGAMENTAL AREA:			
- Epithelial disorders (e.g. carcinoma, papilloma, keratosis) ...	16	2	18
- Reinke's oedema	0	4	4
- Polyps, nodules	8	22	29
- Cysts	2	2	4
- Miscellaneous mild oedema, redness, laryngitis	11	14	26
DISORDERS OF THE CARTILAGINOUS AREA	7	5	13
PALSIES	6	4	11
SUPRA-GLOTTIC LESIONS	1	0	1
TOTAL	51	53	106

Table 6.9 Classification of laryngeal disorders contained in the pathological speaker groups (PM, PF and PA) and number of cases per disorder.

between pathological and healthy speakers is significantly affected by speaker gender.

This evaluation addresses a number of issues related to the assessment of voice samples by acoustic perturbation measures. Firstly, does the normalization procedure for the measurement of F0 and A0 excursions (as explained in Section 5.3 above) successfully account for the general between-speaker differences in F0 values evidenced by the two sexes? If there are no significant differences between groups based on gender then it can be assumed that the normalization procedures were successful. However, significant differences for the factor of sex do not necessarily mean that the normalization technique was unsuccessful particularly if there is an actual trend for male speakers to phonate with greater amounts of perturbation as compared to female speakers. Secondly, if sex of the speaker is not a significant factor for the evaluation of F0 and A0 perturbation parameters then the data for the 2 sexes may be pooled together for a given voice condition. The increased sample sizes would be particularly useful in any discrimination procedures used as a screening tool for the general population of speakers.

SECTION 6.4.1 -- STATISTICAL PROCEDURES

A 2-way analysis of variance (ANOVA) was completed for the 10 acoustic measures derived from the voice samples of the speakers in the groups CA and PA to determine the effects of the factors voice condition (VC) and SEX. A factorial design with unequal cell frequencies was used since the subgroups within each main speaker group differed in sample size (see Table 6.7 for the sample sizes). The use of unequal cell sizes results in three possible main effects

including the effect of voice condition, sex and the additive effects of VC plus sex. The classical experimental approach was used for the ANOVAs and therefore it is possible for significant main effects to be found for each of the individual factors without a significant main additive effect.

SECTION 6.4.2 -- RESULTS AND DISCUSSION

The results of the ANOVAs for the 10 acoustic parameters are shown in appendix 1. There are 10 sets of data, one per parameter, which include the sum of squares for each effect, the degrees of freedom, the mean square values and the F statistic used for significance testing. Significance was regarded as reached by any effect at a level of $p<.01$. Table 6.10 is a summary of the results for significance testing for each parameter; the table includes the tests for the VC X SEX interaction effect, the additive main effects of VC plus SEX and the individual main effects of VC and SEX.

SECTION 6.4.2.1 -- Intonational parameter F0-AV

It can be seen from Table 6.10 that the parameter F0-AV demonstrated a significant VC X SEX interaction for the CA and PA speaker groups. The effect of the condition SEX is not uniform across the 2 categories of voice condition and therefore speakers should be grouped by both conditions to achieve better discrimination results. The significant interaction for this parameter suggests that it would be useful for discriminating between pathological and healthy speakers when the effect of SEX is controlled. The main effects were not tested for significance due to the significant findings for the interaction effect.

PARAM.	INTERACTION EFFECT OF VCOND X SEX	MAIN ADDITIVE EFFECT OF VCOND + SEX	MAIN EFFECT VCOND	MAIN EFFECT SEX
FOAV	S	-	-	-
FODEV	N	S	N	S
JAVEX	N	S	S	S
JDEVEX	N	N	N	S
JRATEX	N	S	S	S
JDPF	N	S	S	S
SAVEX	N	S	S	S
SDEVEX	N	N	N	N
SRATEX	N	S	S	S
SDPF	N	S	S	S

Table 6.10 Summary of results for analysis of variance for the factors VCOND and SEX. S — significant effect; N -- nonsignificant effect.

SECTION 6.4.2.2 -- Intonational parameter F0-DEV

It was found for this parameter that the interaction effect was not significant. Therefore, the main effects were tested for significance which resulted in a significant finding for the main factor of SEX and a non-significant result for the main factor of VC. It was noted for the significant SEX condition effect that the female speakers in both the CA and PA groups produced F0-DEV averages which are larger than for the male speaker group means. That is, the female speakers produced utterances with a wider range of intonation as compared to the male speakers. If the F0-DEV parameter were to be used in the discrimination of pathological and healthy voices then the speakers would need to be grouped by sex as well as voice condition. The lack of discrimination by this parameter for voice condition suggests it is not a useful measure for screening procedures.

SECTION 6.4.2.3 -- Frequency perturbation parameters J-AVEX, J-DEVEX, J-RATEX and J-DPF

None of the four jitter measures demonstrated significant VC X SEX interaction effects. Therefore, the main effects for each parameter were evaluated for significance. The parameter J-AVEX demonstrated a significant effect for the factor of VC and a non-significant effect for the factor of SEX. J-AVEX is the only parameter of the 10 acoustic measures which could be used to differentiate between pathological and healthy speakers without adjusting for the effects of sex of the speakers. On the other hand, the parameter J-DEVEX appears to be only useful for differentiating between the sexes and not between pathological and

healthy voice conditions. The remaining 2 F0 perturbation parameters, J-RATEX and J-DPF, revealed significant main effects for both VC and SEX of speaker. Therefore, both measures are useful for differentiating between pathological and healthy speakers particularly if sex is also used as a grouping factor.

SECTION 6.4.2.4 -- Amplitude perturbation parameters S-AVEX, S-DEVEX, S-RATEX and S-DPF

As in the case of the jitter measures, none of the shimmer parameters revealed significant VC X SEX interaction effects. However, sex of the speaker is a significant main factor for the parameters S-AVEX, S-RATEX and S-DPF. These 3 shimmer measures also demonstrated significant main effects for the factor of voice condition. Therefore, these parameters would be useful for discriminating between healthy and pathological speakers with the best results occurring when sex of speaker is used as an additional grouping factor. Finally, the parameter S-DEVEX did not achieve significance for any of the experimental effects. This finding is in agreement with the observed non-normal highly skewed sample distributions associated with this parameter for each speaker group. Thus, S-DEVEX would not be a useful parameter for discriminating between pathological and healthy groups of speakers.

SECTION 6.4.2.5 -- Observations of multiple classification analysis results

Some general observations can be drawn from multiple classification tables associated with parameters which were not significant for the VC X SEX interaction. For the perturbation parameters excluding S-DEVEX, it was observed that 1) the separation between group means for the control and pathological groups increased as the factor of SEX was accounted for and 2) the separation between group means for each sex increased as the condition of VC was accounted for. Observation of the tables also supports the previous observations that the control group demonstrated lower mean perturbation values as compared to the pathological group and that the male speaker group means were greater than those of the female speakers.

SECTION 6.4.3 -- Conclusions of ANOVAs

The overall conclusion is that a majority of the intonational and perturbational parameters are significantly different for both factors of voice condition and sex. Therefore, the best discrimination of the control and pathological speakers will be achieved when sex of the speaker is also used as a grouping factor. These findings suggest that either 1) the normalization procedure for the measurement of F0 and A0 excursions is not adequate for complete normalization between sexes of speaker or 2) the male speakers in this study produced voice samples with greater perturbation as compared to the data of the female speakers. The results would suggest that both of these factors contribute to the significant sex effects found for most of the parameters. Firstly, the mean values for each perturbation parameter are fairly close between the 2 sexes when voice condition is accounted for despite

the significant main effects for the factor of SEX. The jitter measure J-AVEX was found to be not significant for the main effect of speaker gender. Secondly, it has been reported in the literature that the voices of healthy male speakers are perceived to be rougher as compared to healthy female speakers.

SECTION 6.5 — WITHIN SPEAKER GROUP CORRELATIONAL ANALYSIS

In the preceding section, it was established that the factor sex has a significant effect for most of the intonational and perturbation parameters extracted from samples of connected speech produced by groups of healthy and pathological speakers. The general conclusion drawn from the ANOVAs is that the best discrimination results between the pathological and control speakers is to be achieved when speaker gender is used as a grouping factor (i.e. control males vs pathological males and control females vs pathological females). In the present section, the correlations between the 10 acoustic measures are evaluated within each of the 4 speaker groups CM, CF, PM and PF. In general, it is expected that positive correlations will be found between parameters which are of similar type, that is, between the parameters within the general categories of F0 intonational, F0 perturbational and A0 perturbational measures. Due to the distributional nature of most of the acoustic parameters, correlational analysis of the first-order type can reveal information about the shapes of the individual samples (as extracted from each speaker's voice sample) particularly for the various perturbation parameters.

SECTION 6.5.1 -- STATISTICAL PROCEDURES

Pearson correlation coefficients were derived for each group of speakers (i.e. groups CM, CF, PM and PF) to determine the type and strength of correlation between all pairings of the 10 acoustic parameters within each group. The output of correlational analysis is the determination of the extent to which variation in one variable is linked to the variation in another. The Pearson correlation analysis is a measure of "goodness of fit" of a linear regression line for any 2 given variables under examination. The results of the analysis ~~is a~~ are measures of association indicating the strength of the linear relationship between the 2 variables. Scatter plots were plotted for all 2 variable combinations of the 10 parameters for each group of speakers — the assumption of a linear relationship appeared to be valid for those pairs of variables which demonstrated correlational behavior.

SECTION 6.5.2 — RESULTS AND DISCUSSION

Appendix 2 presents the Pearson correlation coefficients for the pairings of the 10 acoustic parameters, there are four sets of matrices representing the results of the correlational analysis for each group (A2a - CM, A2b - CF, A2c - PM and A2d - PF). Three observations can be made for each one of the coefficients. Firstly, the type of correlation is denoted by the sign of the coefficient, that is, positive correlations between 2 variables have no sign while negative correlations are represented by a minus sign. Secondly, the strength of the correlation is represented by the magnitude of the correlation coefficient where values of -1.0 and 1.0 represent perfect negative and positive correlations, respectively, and values near zero mean little or no correlation

between variables. In the present study, a HIGH correlation is designated as a coefficient greater than .70 (ignoring the sign of the correlation coefficient), a LOW correlation is designated for coefficient values of less than .25 with the remaining intermediate coefficient values representing MODERATE correlations. Thirdly, a given correlation coefficient is considered a significant result for a test of significance as measured at the $p<.01$ level -- significant correlations are marked by a star (*) in the coefficient matrices. Table 6.11 displays a general summary of the degree and type of correlation which were found to be significant — summarized results are presented for each speaker group in terms of general type of acoustic parameter (e.g. F0 perturbation, A0 perturbation, etc.). Non-significant correlations are not presented in Table 6.11.

SECTION 6.5.2.1 -- Correlations between F0 intonational parameters F0-AV and F0-DEV

Significant, positive correlations were found between F0-AV and F0-DEV for all 4 groups of speakers. That is, increasing average F0 level for a given voice sample is associated with increasing range of F0 values. One possible explanation for this finding is a psychoacoustic one. Speakers who evidence a higher average level of pitch for connected speech must also produce a wider range of pitch in order to elicit the equivalent auditory sensation of pitch range as produced by speakers who phonate with lower average levels of pitch. However, the evidence is not completely supportive of this explanation as only moderate correlations between F0-AV and F0-DEV were found for the female groups as compared to the more supportive high correlations of the male speaker groups. Further investigation

PARAMS.	CM	CF	PM	PF
FOAV/ FODEV	HIGH/POS	MOD/POS	HIGH/POS	MOD/POS
JITTER	MOD-HIGH/POS	HIGH/POS	HIGH/POS	HIGH/POS
SHIMMER	MOD-HIGH/POS	MOD-HIGH/POS	HIGH/POS	MOD-HIGH/POS
JITTER/ SHIMMER	MOD/POS	HIGH/POS	MOD/POS	MOD-HIGH/POS
FOAV/ JITTER	---	MOD/NEG	---	MOD/NEG
FOAV/ SHIMMER	MOD/NEG	MOD/NEG	MOD/NEG	MOD/NEG
FODEV/ JITTER	---	---	MOD/POS	---
FODEV/ SHIMMER	---	---	---	---

Table 6.11 Summary of general correlation results using Pearson correlation coefficients. MOD -- moderate; NEG -- negative; POS -- positive.

of this sex difference for F0-AV/F0-DEV correlations may best be achieved by an analysis of covariance in which the effects of speaker gender is controlled for as a nuisance factor.

SECTION 6.5.2.2 — Correlations between F0 perturbation parameters J-AVEX, J-DEVEX, J-RATEX and J-DPF

For all 4 speaker groups, significant positive correlations were revealed between the 4 parameters of frequency perturbation. These results support the expectation that measures of perturbation co-vary such that increasing levels of F0 waveform perturbation are reflected as increases in all 4 acoustic measures. The degree of the correlations between the F0 perturbation parameters is related to the nature of the original F0 excursion sample distribution extracted for the voice sample produced by each speaker. High correlations are expected between J-AVEX and J-DEVEX if the signed F0 excursions extracted from a given voice sample are distributed in a normal fashion about a zero mean excursion value (i.e. signed excursions are normally-distributed about the smoothed trend line of F0). J-DEVEX is the standard deviation of the distribution of signed excursions. J-AVEX is derived from the equivalent magnitude excursions from the smoothed trend line which are distributed with a large number of excursions toward the lower end of the excursion values and skewed in a positive direction towards the higher end. If the signed excursions are distributed normally then the mean of the magnitude excursions distribution is determined by the standard deviation of the signed excursions distribution. That is, J-AVEX may be interpreted as reflecting the magnitude of the dispersion of the signed excursions (this is an extension of the logic first set

out by Horii 1979 for cycle-to-cycle measures of average jitter). In this case, J-AVEX and J-DEVEX are somewhat redundant measures of frequency excursion distributional range and J-AVEX should not really be interpreted as a measure of central tendency (the median value of the magnitude excursions may be a more appropriate alternative). High correlations are in fact found between J-AVEX and J-DEVEX for the 3 speaker groups CF, PM and PF. J-RATEX is also highly correlated with J-AVEX and J-DEVEX for these 3 groups — the degree of these correlations is expected since J-RATEX measures the number of occurrences of excursions in the tail regions of the distribution of greater than 3% (i.e. the number of occurrences in both tails of the signed excursions distribution or in the single tail of the distribution of the magnitude excursions). A moderate correlation was found for the speaker group CM between parameters J-AVEX and J-DEVEX which suggests less than normal distributions of the excursion samples extracted from each control male speaker's recorded utterance. This finding makes the interpretation of the measures J-AVEX and J-DEVEX more difficult though J-AVEX is still highly correlated with J-RATEX for this group. As a measure of directional changes in an unsmoothed F0 contour, J-DPF is positively correlated with the other 3 perturbation measures. That is, the percentage of substantial directional changes in a given F0 contour increases as the general range of the excursions for the contour also increases. For all 4 speaker groups, a very high correlation is evidenced between J-RATEX and J-DPF and therefore these 2 measures provide somewhat redundant information about F0 perturbation behavior in a given voice sample.

SECTION 6.5.2.3 -- Correlations between A0 perturbation parameters

S-AVEX, S-DEVEX, S-RATEX and S-DPF

For most of the variable pairings, significant positive correlations were revealed in each of the 4 speaker groups. This is further evidence for the tendency for the measures of perturbation derived from samples of connected speech to co-vary. The moderate strength of the correlations of the S-AVEX/S-DEVEX pairing for 3 of the groups (i.e. CM, CF and PF) would suggest that the distributions of the signed amplitude excursions about the equivalent smoothed trend lines were not normal for many of the speakers within each of these groups. The interpretation of the S-AVEX and S-DEVEX is therefore difficult since neither measure fully explains the distributional range behavior of a given sample of amplitude excursion values. The unusual distributional characteristics of a given sample of A0 excursions is further suggested by the lack of significant correlations between S-DEVEX and S-RATEX for the PM group even in the presence of a high correlation between S-AVEX and S-DEVEX. A moderate or high correlation was demonstrated between S-AVEX and S-RATEX for the 4 groups which suggests that S-AVEX would be a more useful indicator of A0 perturbation than S-DEVEX. The parameter S-DPF revealed positive correlations with the S-AVEX and S-RATEX measures for all the speaker groups. That is, increased numbers of substantial directional changes in unsmoothed A0 contours are associated with increased measures of range of amplitude excursions about their equivalent smoothed A0 contours. The strengths of the correlations between the parameters S-RATEX and S-DPF are high and therefore these A0 perturbation measures may be somewhat redundant.

SECTION 6.5.2.4 -- Correlations between F0 and A0 perturbation parameters

In general, most F0 perturbation parameters are positively correlated with the A0 perturbation measures for the 4 speaker groups. These findings are not unexpected since both types of perturbation measures examine irregularities in contours which reflect vibrational patterns in samples of connected speech. The tendency for moderate levels of correlations between F0 and A0 perturbation suggests that the 2 types of perturbation measures are not overly redundant in their information (although there are quite a few high correlations to be seen between these parameters in the speaker groups). One notable exception is the parameter S-DEVEX which for the most part is not significantly correlated with the F0 perturbation parameters in 3 of the speaker groups. Further statistical analysis, for example by partial correlations, would be required to tease out the complex and possibly spurious relationships that exist between the various F0 and A0 perturbation parameters. In this future research, it would be useful to determine the degree to which amplitude measurement, used for both pitch period detection and amplitude perturbation measurement, affects F0 perturbation analysis.

SECTION 6.5.2.5 -- Correlations between F0 intonational and perturbation parameters

From the correlational analysis, it was revealed that significant negative correlations exist between the parameter F0-AV and many of the F0 and A0 perturbation parameters. Moderate negative correlations were noted between F0-AV and the measures of

jitter for the 2 female speaker groups while no significant correlations were demonstrated for the male groups for these measures. In the female case, it is suggested that increasing average F0 is moderately associated with decreasing values of F0 perturbation. A possible explanation for the negative relationship between F0-AV and jitter is that higher average levels of F0 are associated with increased tension in the vocal folds during vibration — this increased tension is also linked with increasing efficiency in the vibration of the vocal folds. However, the evidence supporting this explanation is not very substantial since the correlations are at best moderate for the female groups and non-existent for the 2 male groups. For all 4 speaker groups, significant negative correlations were revealed between F0-AV and the measures of A0 perturbation (except for the measure S-DEVEX). It would appear that increasing average F0 is associated with decreasing levels of amplitude perturbation as evidenced in samples of connected speech. The role of tension factors in the vibration of the vocal folds is also suggested by the moderate correlations between F0-AV and shimmer measures. Further research is required to determine the sources of the correlations between F0-AV and the perturbation measures. Detailed correlational analysis of the sample values which determine F0-AV and perturbation measures for each speaker's voice sample may reveal whether the correlations between these parameters are due to actual physiological causes and/or systematic ones (e.g. the correlations are due to the mathematical calculations of the normalization procedures, sampling resolution, pitch extraction errors, etc.). The finding of a sex difference for F0-AV/F0 perturbation parameter correlations may be clarified by an analysis of covariance in which adjustments could be

made for the effects of speaker gender. The results for the analysis of F0-DEV and the perturbation parameters suggest that, for the most part, the 2 types of parameters are not significantly correlated.

SECTION 6.5.3 -- CONCLUSIONS OF CORRELATIONAL ANALYSIS

Positive moderate-to-high correlations are found between each type of acoustic parameter (i.e. F0 intonational, F0 perturbational and A0 perturbational measures) for each of the speaker groups. Positive correlations are noted across parameter type for the F0/A0 perturbation correlational analyses. The evaluation of F0 intonational/perturbational relationships demonstrated moderate negative correlations between these parameters. The results would suggest that many of the parameters represent redundant information (e.g. very high correlations between various F0 perturbation measures) which should be accounted for when a multivariate statistical procedure is used to discriminate between the groups of pathological and control speakers. In addition, the correlational analyses were of the first-order type -- detailed partial correlational analysis could reveal more information about the relationships between the various parameters. The effects of the factor sex of the speaker may have biased the results of the correlational analysis and therefore further analysis of covariance of the data would be useful. Finally, the distributional nature of most of the acoustic parameters suggests that detailed analysis of the sample distributions derived from each speaker's voice sample should be completed in order that we might better understand the systematic sources of variation in the group data.

SECTION 6.6 -- DETECTION OF PATHOLOGICAL AND CONTROL SPEAKERS BY PATTERN RECOGNITION TECHNIQUES

It was previously demonstrated by analysis of variance that significant differences exist between the pathological and control groups of speakers for a number of the intonation and perturbation parameters. In this section, a pattern recognition technique is used to combine the discrimination power of the individual parameters into a multivariate statistical procedure for classifying pathological and control speakers. That is, the pattern recognition technique is applied to the acoustic data to evaluate its usefulness as a tool for screening speakers for pathological conditions of the voice. This section begins with a brief description of the pattern recognition technique based on the Maximum Likelihood principle -- the results of applying this classification technique to the pathological and control speakers' data is presented in the remainder of this section.

SECTION 6.6.1 -- PATTERN RECOGNITION USING THE MAXIMUM LIKELIHOOD PRINCIPLE

Pattern recognition techniques have been used in other studies in order to discriminate between healthy and pathological speakers (see, for example, Murray and Doherty 1980 and Zyski et al. 1984 in which discriminant analysis was used for detection purposes). The following discussion of the Maximum Likelihood classification method is essentially a non-mathematical and practical one. Detailed discussions of the theoretical and mathematical issues behind pattern recognition techniques can be found in the works of Nagy (1968), Duda and Hart (1973) and Kanal (1974) amongst others. Of

particular relevance is the study completed by Davis (1976; 1979) in which the Maximum Likelihood classifier was applied to a number of acoustic parameters in order to classify speakers as pathological or normal for phonation behavior. The algorithms used in the present study for pattern recognition were also used by Davis -- these programs are widely available as part of a signal processing software package (ILS -- Interactive Laboratory System) produced by Signal Technology Inc.

One of the major objectives of pattern recognition is the design of a classifier which is capable of separating a given set of data samples into one of several categories (Davis 1976). Figure 6.28 (based on Davis 1976 and Kanal 1974) presents the general feature-extraction classification model of pattern recognition (Figure 6.28a) and the specific system as used for classification of speakers in the present study (Figure 6.28b). The measurement process and feature extraction portions of the system have been presented in the chapters on pitch extraction and perturbation measurement algorithms (Chapters 3 and 5, respectively). The classifier itself is constructed from the set of acoustic parameters (i.e. the features) which have been extracted from measurements of the input voice samples. The issues relevant to pattern recognition are the type of classifier, the empirical measurement of the success of the classifier and the method by which the classifier is applied to the test data.

The classifier to be applied to the acoustic data in the present study is based on the Maximum Likelihood principle. This classifier is a Bayesian statistic in which the distributions of the features within each category of the classification (i.e.

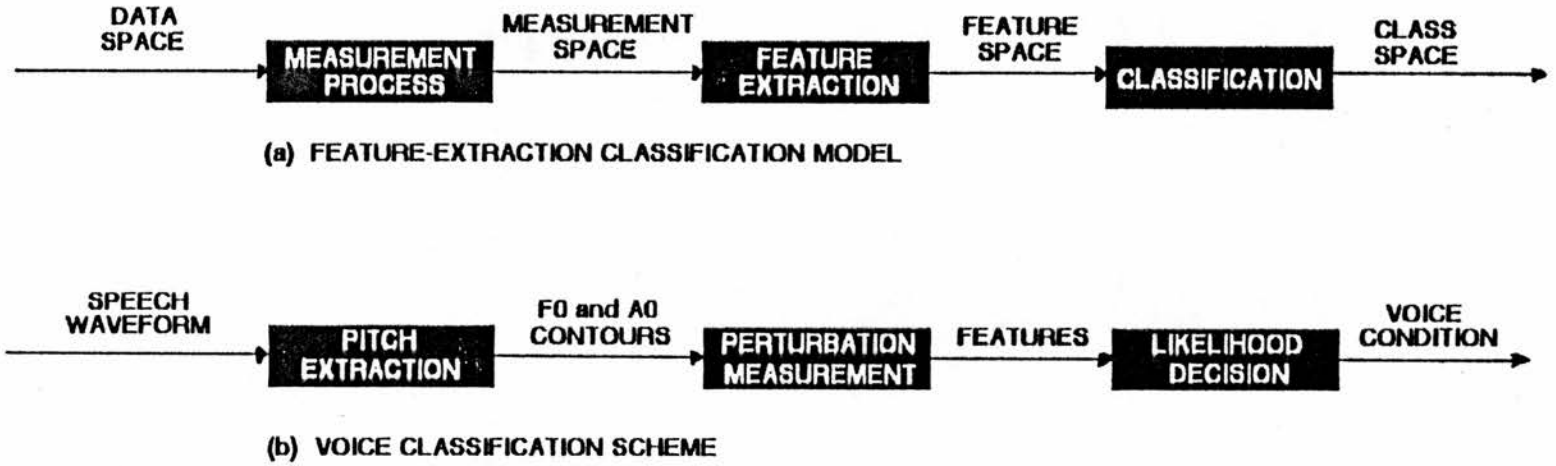


Figure 6.28 (a) The general model for feature extraction used for pattern classification and (b) the specific model used for the classification of voice condition based on the intonation and perturbation parameters. (Figure after Davis 1976)

pathological or control) are represented by probability density functions with gaussian distributional characteristics. By way of example, the Maximum Likelihood principle is first described for the simplest classification task, that is, the use of one feature to classify data into one of 2 possible categories. In the analysis stage, reference groups of data from each category are statistically evaluated to form probability distributions for the single feature, one density function per reference group. These 2 distributions are Bayesian in character with assumed gaussian distributional properties based on the means and variances of the sample distributions of the feature within each reference group. The most useful feature in a discrimination task is one which produces density functions which are well separated with no overlap between categories. In reality, a certain amount of overlap often occurs between the probability distributions which can lead to misclassification of individual cases during the classification procedure. In the classification stage, new test cases (i.e. a feature representing the new case) are introduced and distance measures used to determine the probability of a given new test case being from one or the other of the 2 density functions. The distance measure of the Maximum Likelihood technique also assumes a gaussian criterion for the distributions associated with the 2 reference groups. The multivariate classification task is an expansion of both the number of features and categories to be used for classifying the data. As described above, probability density functions are obtained for each one of the features based on the multiple reference groups. A mathematical procedure is used to collapse the resultant multidimensional space into a single dimension by a linear combination of the probability densities. The

classification of test cases is then based on the new single dimension. The Maximum Likelihood principle includes assumptions about the population distributions from which the sample probability density functions are formed. Firstly, it is assumed that the population from which the features are derived is Guassian in distribution (though the Maximum Likelihood technique is considered quite robust to violations of this assumption). The results of the chi-square analysis of the sample distributions for each feature (see Section 6.3 above) suggest that this assumption is fulfilled for both categories of speakers. Secondly, the a priori assumption is made that the population distributions for the various categories demonstrate equal variances. The standard deviations displayed in Table 6.8 for the features of each group do not actually support the assumption of equal variances, rather the variances appear to be in an approximate ratio of 2:1, pathological versus control.

The usefulness of a given classifier (and therefore of the reference features used to construct the classifier) is determined from empirical estimates of the error probability produced by classification of test cases. Three types of error probability which may be used to evaluate the results of classification include (Davis 1976):

1) Probability of False Alarms — For the present study, this would be the probability that a control speaker is classified as a pathological speaker (also known as a False Acceptance or Type I error).

2) Probability of a Miss — This is the probability of a failure to identify a pathological speaker as being pathological (also known as a False Rejection or Type II error).

3) Probability of Correct Detection — In the present study, this is the probability of correct acceptance in either of the 2 categories. An overall average probability of correct detection can be computed for

the 2 categories together.

Davis (1976) noted that empirical estimates of error probability for a given classifier are reasonable under certain conditions. Firstly, the number of samples in the original data space should be several times larger than the number of features. For the two class problem (i.e. pathological versus control), the number of samples divided by the number of features should be greater than 3. Secondly, the ratio of sample size to feature size should be greater than 10 if the parameters of the multivariate normal distributions (e.g. the means and standard deviations) also must be determined from the input sample data space.

Determination of the empirical estimates of error for a given classifier may be derived in several ways depending on the number of samples within the reference and test groups. The following classification tasks are described by Davis (1976):

1) Leave-One-Out Method -- Given a sample space of N cases, the test group consists of one case and a reference group of N-1 cases is used to construct the classifier. An empirical estimate of the error probability is then determined for the one test case. This procedure is repeated N times with a new test case being from the reference group for each classification (and the previous test cases added back to the reference group). An average error probability for all possible combinations of test and reference groups is calculated for this technique.

2) Hold-Out Method -- Given a sample space of N cases, the cases are split evenly between reference and test groups within each category of classification. The estimate of error probability is determined for the test group and then the reference and test groups are swapped. An average error of probability is then derived based on the results of the 2 original error estimates.

3) Rotation Method -- The size of the test group is greater than one and less than half of the total data space within each category of classification. Probabilities of error are estimated for all possible combinations of test and reference groups and averaged.

The rotation method is a compromise between the leave-one-out and hold-out methods.

When the estimates of error are determined by any one of the above methods and the reference and test groups contain data from different sets of speakers then an upper bound for future correct classification of new cases is represented by the error probability. If the cases within the test groups are the same as those within the reference groups then the task is known as Testing-On-The-Training-Set or Resubstitution. Resubstitution provides a lower bound for future classification of new cases.

In summary, the Maximum Likelihood principle is one type of pattern recognition classifier for separating samples of a data space into several categories. The classifier is based on probability density functions derived for a number of features within each category. The usefulness of the classifier is based on estimates of the error probability for correct classification of test cases based on a given classification procedure.

Two issues remain to be discussed in this section. Firstly, given a large set of features for each sample in the data space, what is the subset of these features which yields the lowest estimates of error for the classifier? The determination of this subset of features will provide the most computationally efficient classifier, lower probabilities of error and highlight the most useful features for classification. Secondly, given the resultant subset of features, what is the probability of error for classifying test groups of pathological and control speakers based on reference groups of these speakers? The following sections present experimental results which address these 2 issues.

SECTION 6.6.2 -- EXPERIMENT I: FORWARD SEQUENTIAL SELECTION OF FEATURES FOR CLASSIFICATION

In this section, an experiment is presented in which the "best" subset of features is derived from the 9 available acoustic parameters including FO-AV, FO-DEV, J-DEVEX, J-AVEX, S-AVEX, J-RATEX, S-RATEX and S-DPF. Given any set of L possible features of N samples, there exists a $K \leq L$ subset of features which will yield the lowest estimate of error probability for a given classifier (Davis 1976). The best subset of features is obtained by evaluating all combinations of the L features, taken K at a time where K is set to 1,2,...,L. One simple computational procedure for determining the best subset is the "Forward Sequential" method. The forward sequential method begins by evaluating each feature singularly to determine the best parameter for classification amongst all the features. The best pair of features is then determined and this pair includes the initial single best feature. More and more features are added by this process until further increases in the number of features no longer decreases the error probability.

The forward sequential method was applied to the features of the pathological and control speakers to determine the best subset of features for classification. This classification task was completed for each sex, that is, speaker groups PM versus CM and PF versus CF. The Maximum Likelihood principle was used as the classifier in each step of the forward sequential analysis. The probability of correct detection for each test group (i.e. pathological and control) was used as the empirical measure of error. The decision criterion for the classifier assumed uniform probabilities of detection of either category of voice condition.

The design of the classifier was based on the resubstitution method in which the reference groups were also used as the test groups.

SECTION 6.6.2.1 -- Results of single feature tests

Figure 6.29 presents the results of the single feature classification tests for the male groups (Figure 6.29a) and the female groups (Figure 6.29b) of speakers. The abscissa of each plot identifies the 9 single features and the ordinate is the probability of correct detection for each feature as found in the resubstitution task. In each part of Figure 6.29, results are presented by test group category (P=pathological, C=control) as well as by average probability of correct detection for both test groups (B=both).

The results of the single feature classification tasks for the male speaker groups are presented in Figure 6.29a. For the control male speakers, the probabilities of correct detection range from 83.33 to 93.93%. A range of correct detection from 35.29 to 90.20% is seen for the male pathological speakers. In general, most of the acoustic features are better at identifying the male control speakers than the male pathological speakers. For both groups of male speakers, the parameter S-DPF produced the highest levels of correct detection.

Figure 6.29b displays the results of the single feature classification tasks as completed for the female speaker groups. The results are similar to the findings of the male speaker groups. The probabilities of correct detection ranged from 79.31 to 91.38% for the female control speakers while a range of 48.08 to 76.92% was found for the female pathological speakers. As in the male case,

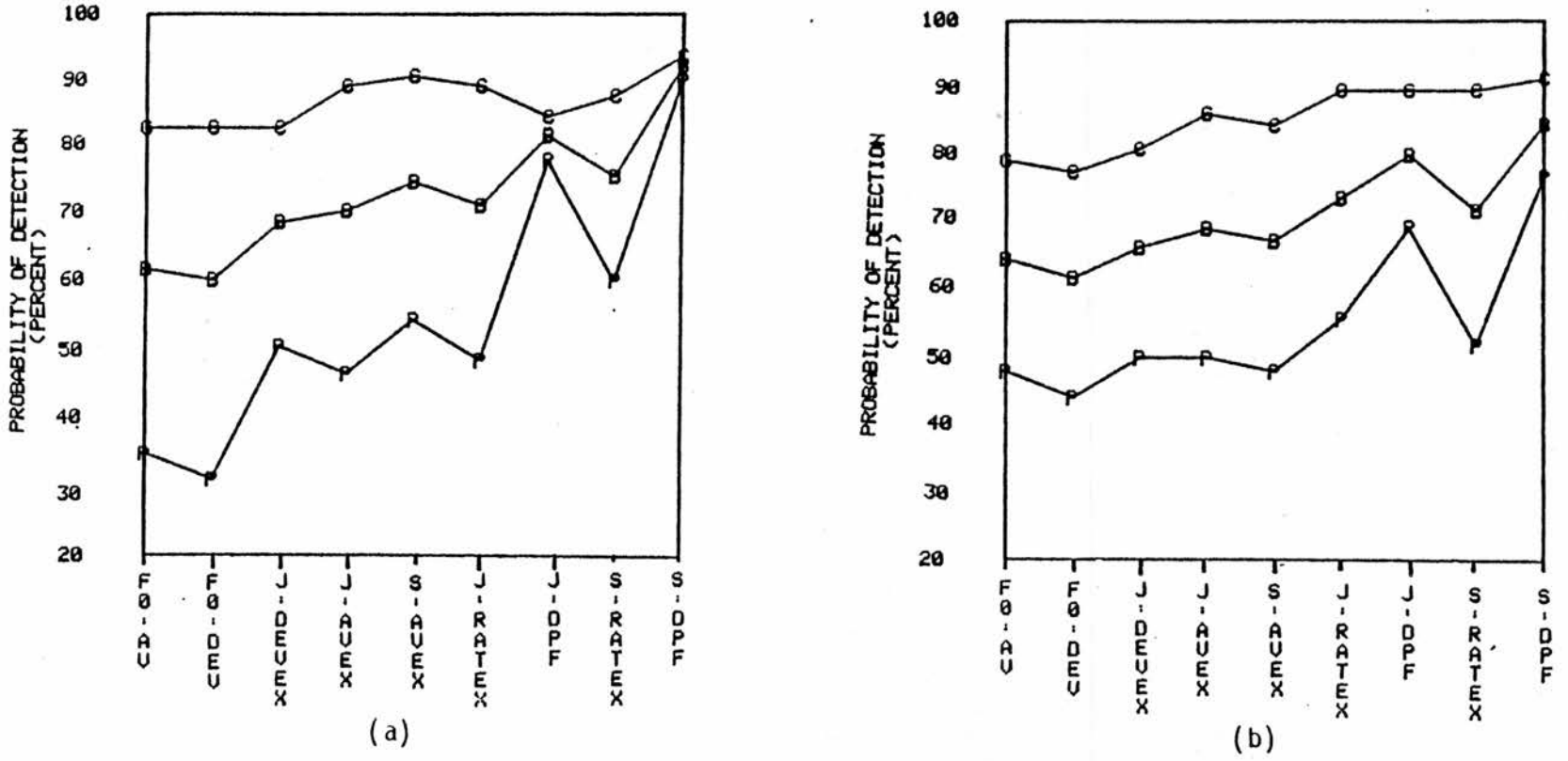


Figure 6.29 Results of single feature classification tasks with the probability of correction detection plotted against 9 features of intonation and perturbation. (a) male speaker groups; (b) female speaker groups. KEY: C -- Control group; P -- Pathological group; B -- average probability for both control and pathological groups.

the single features provide better identification of the control speakers as compared to the pathological speakers. It can also be seen from Figure 6.29b that the feature S-DPF is the best single feature of the 9 available parameters.

Two notable differences are to be found between the sexes where the correct detection rates of the F0-AV and F0-DEV parameters are lower for the male pathological speakers as compared to the female pathological speakers.

SECTION 6.6.2.2 -- Final results of forward sequential classification tasks

For the male groups of speakers, the final average probability of correct detection was found to be 95.73%. Three features were required to reach this level of detection including S-DPF, F0-AV and J-RATEX. This is a satisfactory result in that the best subset of features for the male speaker groups includes a parameter from each type of acoustic measurement (i.e. shimmer, intonation and jitter). It should be noted that the subset of S-DPF, F0-AV and J-AVEX also produced a 95.73% probability of correct detection. The parameter J-RATEX was selected for practical considerations in that it is a straightforward count of the number of large excursions of F0 for a given voice sample. The J-AVEX parameter may be difficult to interpret if it is derived from a distribution with non-normal properties. The 3-feature subset of S-DPF, F0-AV and J-RATEX produced a greater correct detection rate than any of the single features as well as compared to all 9 features together (probability of correct detection equal to 93.16%). For the female speaker groups, a 6-feature subset including S-DPF, J-AVEX, J-RATEX, F0-AV,

F0-DEV and S-RATEX produced an overall probability of correct detection equal to 97.27%. This is a greater level of correct detection than found for any of the single features or for a set of 9 parameters (probability of correct detection equal to 95.45%). As in the male case, other combinations of features produced equivalent levels of correct detection for the female speaker groups -- the final 6-feature subset was determined by practical considerations such that 2 features representing each type of parameter are included in the subset (the distributional types of perturbation measure were left out where possible). The number of features contained within each best subset are well within the limits required to produce reasonable empirical estimates of error probabilities for the speaker groups.

SECTION 6.6.3 -- EXPERIMENT II: OPEN TEST CLASSIFICATION OF PATHOLOGICAL AND CONTROL SPEAKERS

In this section, results are presented from an experiment in which speakers are classified as pathological or control by the Maximum Likelihood classifier in an open test paradigm. The purpose of these classification tasks is to determine its usefulness as a tool for screening speakers who evidence pathological conditions of the voice.

Two sets of open classification tests were completed, one set per gender. The Maximum Likelihood classifier was designed and evaluated using the hold-out method. This design requires reference and test groups of equal sample sizes for each category of classification. A randomization procedure (i.e. coin tossing) was used to split the 4 speaker groups into reference and test groups.

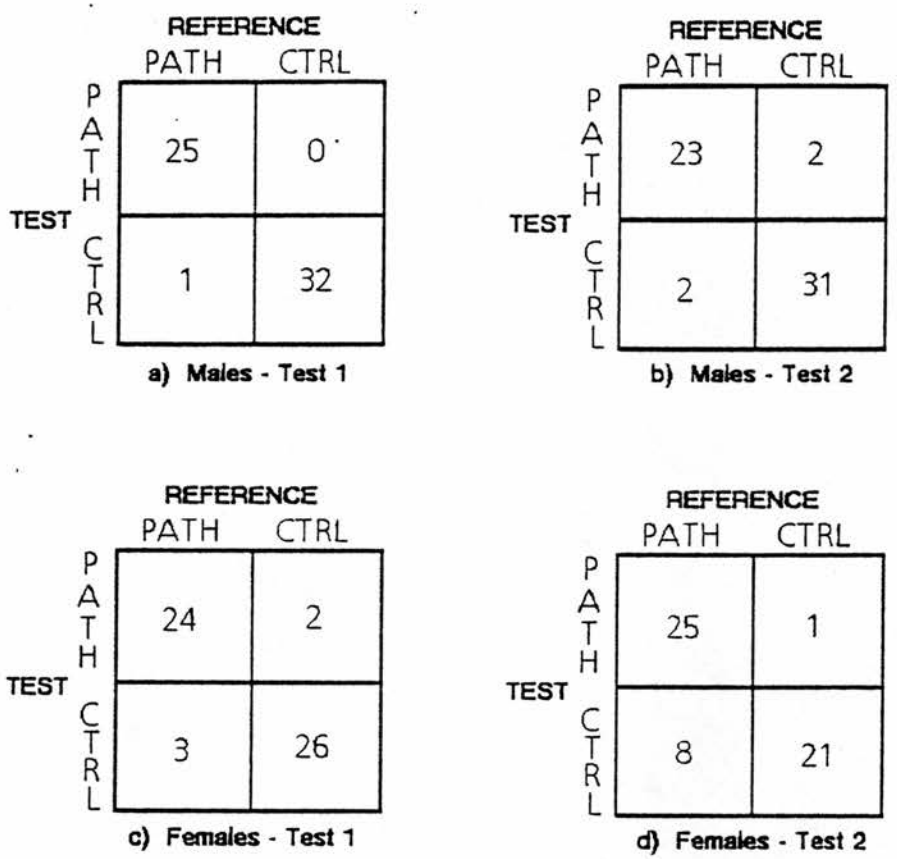


Figure 6.30 Confusion matrices presenting the results of the open test classification tasks. a) Test 1 of hold-out method for the male speaker groups; b) Test 2 for the male speaker groups; c) Test 1 of hold-out method for the female speaker groups; d) Test 2 for the female speaker groups.

TEST GROUP	CORRECT ACCEPTANCE PATHOLOGICAL	CORRECT ACCEPTANCE CONTROL	AVERAGE ACCEPTANCE BOTH VCOND	FALSE ALARMS
MALE TEST 1	100.00	96.97	98.28	3.03
MALE TEST 2	92.00	93.94	93.10	6.06
MALE AVE.	96.00	95.46	95.69	4.54
FEMALE TEST 1	92.31	96.15	94.23	3.85
FEMALE TEST 2	89.66	72.41	81.04	27.59
FEMALE AVE.	90.91	83.64	86.65	16.36

Table 6.12 Results of pattern recognition classification of pathological and control speakers (scores are listed in percentages).

For the male speaker classification tasks, the CM group was split into reference and test groups of 33 speakers each while the PM reference and test groups consisted of 25 speakers apiece (the one extra PM speaker was thrown out of the experiment). For the female speaker classification tests, there were 29 speakers per reference and test groups for the CF group and the PF reference and test groups each contained 27 speakers. For each gender, a classifier was created based on the pathological and control reference groups using the best subset of features which were previously determined for each sex. The classifiers were constructed with equal a priori probabilities for detecting either the pathological or control voice conditions. The classifier was then applied to the test groups to determine empirical estimates of the error probabilities. The samples within the test and reference groups were then switched for each category of voice condition -- the construction and evaluation of the second half of the hold-out method of classification was then completed for each gender.

The results of the classification tasks are presented in Figure 6.30 and Table 6.12. In Figure 6.30, a confusion matrix is presented for each classification task. There are 2 matrices per gender which represent the results for the 2 halves of the hold-out classification method. In each matrix, the reference groups are represented in the columns and the test groups are represented by the rows. The upper right portion of each matrix displays the number of correct acceptances for a given pathological test group while the upper left portion presents the number of misses for that pathological group. The number of correct acceptances for a given control test group is shown in the lower right portion of each

matrix while the number of misses for the control group is seen in the lower left portion.

Table 6.12 presents the information contained in the various matrices in terms of percentage error probabilities. This table includes information for the percentage of correct acceptance for both the pathological and control test groups. In addition, the percentage of false alarms is displayed in Table 6.12 -- a false alarm is defined as a control speaker who has been misclassified as a pathological speaker. The percentage of correct acceptance for each pathological group can also be viewed as the percentage of correct detection of the pathological voice conditions and used in conjunction with the percentage of false alarms. The error probabilities are presented for male and female groups of speakers -- there are 2 trials per gender due to the hold-out method plus the average results across the 2 trials.

It can be seen from Table 6.12 that the classification of the male speakers resulted in high percentages of correct acceptance (92% or better) for both categories of voice condition, in both trials of the hold-out method. The overall average percentage of correct acceptance for both pathological and control male speakers was 95.69%. The percentage of correct detection of male pathological speakers was found to be 96.00% with a false alarm rate of 4.54% for misclassified control speakers. For the female speakers, slightly lower percentages of correct acceptance were found as compared to the male speakers but the percentages are still considered to be at least moderate (72.41% or better). The overall average percentage of correct acceptance for both pathological and control female speakers was 86.65%. The average percentage of

correct detection for the female pathological speakers was 90.01% with an average false alarm rate of 16.36%. The average scores for the female speakers have been lowered primarily by the percentage of false alarms found in the second trial of the hold-out method (i.e. 8 of 29 control females were misclassified). The percentages of correct acceptance of the pathological speakers are comparable between the 2 genders. The percentages of correct detection found in the present study are higher than the results presented by Davis (1976). In an open test using the Maximum Likelihood classifier constructed from 6 acoustic voice features, Davis revealed a percentage of correct detection of pathological speakers equal to 67.4% and a false alarm rate of 21.7% (the speakers were not separated into groups according to gender).

Thus, it would appear that future screening of the population for pathological conditions of the voice is scientifically feasible through the voice analysis and pattern classification schemes described above. The resultant scores of the classification tasks were quite high with relatively low false alarm rates for most of the groups. It is interesting to note that these high detection rates were achieved with no attempt to bias the classifier towards higher levels of correct detection of the pathological speakers (i.e. 100% correct detection of pathological voices). It is presumed that even higher levels of correct detection of pathological voices can be achieved without a large increase in the false alarm rate (though this issue has not been specifically tested in the present study). However, the results of the classification tasks should be treated with some caution. Firstly, and most importantly, the high correct detection rates and low false alarm

rates are due to a certain extent to the original speaker selection procedures which eliminated speakers with feature values that were considered unrepresentative of their group as a whole. Normal sample distributions were achieved for each feature and therefore the variances were artificially trimmed for each voice condition category. Secondly, Duda and Hart (1973) and Davis (1976) observed that the empirical estimates for the error probability should be treated with caution unless the number of samples used in the classification task is large. The number of speakers used in the present study is higher on average than most studies of voice perturbation but still considered low for a statistical technique such as pattern recognition. Given the low speaker numbers, it is possible to have higher probabilities of error than ~~has~~ have been displayed for the particular sets of test groups used in the present study. Thirdly, the classification tasks presented in this study are not true examples of screening since the pathological groups consisted of speakers who evidenced diagnosed disorders of the vocal folds (i.e. there was a priori knowledge of these speakers' voice conditions). However, the results do suggest a reasonable platform from which the perturbation measurement system may be applied to unknown states of the vocal folds prior to medical diagnosis.

SECTION 6.7 — SUMMARY

A series of experiments has been presented in this chapter which investigated the use of the perturbation measurement system for analyzing samples of connected speech produced by healthy and pathological speakers. Two experiments investigated the nature of the recorded speech sample which is to be input to the perturbation

measurement system. In a durational study of the input speech sample, it was shown that approximately 40 seconds of read connected speech is required from a speaker in order for long-term measurements of the acoustic intonational and perturbational parameters to stabilize and characterize the phonatory performance of male and female healthy speakers. In an investigation of the effects of speech signal phase compensation, it was found that low-frequency phase compensation of tape recorder-induced distortions of voice samples resulted in greater number of significant differences between a healthy and pathological group of speakers for the various acoustic parameters. It is concluded that 40 seconds of phase compensated (or undistorted) read connected speech is an adequate input signal to the perturbation measurement system.

A further four statistical procedures were completed to determine the ability of the parametric outputs of the perturbation measurement system to distinguish between groups of pathological and control speakers. In the first procedure, speakers were eliminated from available pools of pathological and control subjects if a number of their associated acoustic parameters were found to be highly unrepresentative of their respective groups as a whole. Chi-square analyses of the resultant speaker groups (i.e. control all, control male, control female, pathological all, pathological male and pathological female) revealed sample distributions for most of the acoustic parameters which were considered representative of statistically normal population distributions. It is concluded that these speaker groups were suitable for the remaining parametric statistical analyses. The second procedure used analysis of

variance of the control and pathological groups to determine the effects of the features VOICE CONDITION and SEX on the discrimination of the speakers by the acoustic parameters. In general, it was found that while CONDITION X SEX interactions were not significant, the 2 main effects of VOICE CONDITION and SEX were significant for most of the acoustic parameters. It was concluded that the differentiation of pathological and control speakers is best realized when the speakers are also grouped for gender as well as condition of voice. The analysis of variance demonstrated that, in general, the pathological group of speakers produced significantly greater levels of perturbation as compared to the control groups. The third statistical procedure examined the first-order correlations between the various acoustic parameters within each speaker group. In general, high positive correlations were found within each broad type of acoustic parameter (i.e. F0 intonational, F0 perturbational and A0 perturbational measures) for all groups of speakers. Moderate-to-high positive correlations were revealed between F0 and A0 perturbational parameters while moderate negative correlations were demonstrated between intonational and perturbational parameters. It is concluded that redundant information regarding phonatory efficiency was displayed by many of the acoustic parameters. In addition, there is a general trend for lower perturbation values as the average level of F0 increases. In the final set of experiments, the pathological and control speakers were classified into groups by the Maximum Likelihood pattern recognition technique. The "best" subset of features amongst the available 9 acoustic parameters for classifying pathological and control speakers was determined for each gender by resubstitution classification tasks using a forward sequential method. For the

male speakers, a good lower bound of correct classification is achieved by a 3-feature subset consisting of the parameters S-DPF, FO-AV and J-RATEX. A 6-feature subset including S-DPF, J-AVEX, J-RATEX, FO-AV, FO-DEV and S-RATEX is required to produce a good lower bound of correct classification for the female speakers. For both genders, high levels of correct detection of pathological speakers and low levels of false alarms for misclassification of control speakers were demonstrated by open classifications of control and pathological speakers using the best subsets of features. It is concluded that a useful system has been created for distinguishing between pathological and healthy speakers based on acoustic measures of intonation and perturbation derived from samples of connected speech.

CHAPTER 7

CONCLUSION

CHAPTER 7

CONCLUSION

This concluding chapter begins with a summary of the main points presented in the previous chapters of this thesis. Future areas of research are then discussed for the medium-term development of the perturbation measurement system as well as its future application to the analysis of healthy and pathological phonation.

In this thesis, a system was developed for the acoustic analysis of waveform perturbations displayed in samples of connected speech. The system was designed to provide objective quantitative measures of signal irregularities which characterize detailed fundamental frequency and amplitude contours extracted from speech waveforms. The successful development and evaluation of the system, as presented in Chapters 2 to 6 above, are the initial steps towards its future application to three primary objectives of laryngeal pathology research, that is, screening, diagnostic support and long-term monitoring of laryngeal pathologies. The perturbation measurement system consists of three major components including: 1) primary extraction of F0 and A0 contours from a sample of connected speech, 2) smoothing of the two contours to produce trend lines from which waveform perturbations can be measured and 3) statistical evaluation of long-term parameters of intonation and perturbation. The system consists of computer programs written in FORTRAN and operates on a VAX 11/750 minicomputer (a PDP 11/40 minicomputer is also used to digitize the speech samples).

In Chapter 2, a broad typology of pitch detection algorithms was presented, and a number of major types of PDA was exemplified and discussed. A multichannel solution based on temporal structural analysis appears to be the best choice of time domain PDA in order to process the wide variety of signals produced during connected speech, particularly by speakers with pathological laryngeal structures. In Chapter 3, a modified version of a parallel processor operating in the time domain (Gold and Rabiner 1969) was implemented as the primary component for extracting F0 and A0 contours for perturbation analysis. The parallel processor is considered to be a powerful but relatively uncomplicated time domain PDA due to its multichannel approach to structural analysis of the speech waveform. The necessary conditions for the extraction of useful pitch data for perturbation analysis were also discussed in detail in Chapter 3. A literature review of the previous investigations into the perturbatory behavior of speech waveforms was presented in Chapter 4. The majority of these perturbation studies have measured cycle-to-cycle parameters of frequency perturbation in samples of sustained vowel phonations produced by healthy and pathological speakers. Only a few studies have completed perturbation analysis on samples of connected speech and, in particular, using a trend line approach to evaluate frequency and amplitude perturbations. In Chapter 5, the perturbation measurement algorithms used in the system were described in detail. The notable feature of this component of the system is the non-linear smoothing of the contours which produces trend lines of F0 and A0, from which excursions of the unsmoothed input values may be estimated. The non-linear smoother chosen for this purpose consists of a 5-point running-median plus a 3-point Hanning window, a system first

described by Rabiner et al. (1975). A series of experiments were presented in Chapter 6 which evaluated the performance of the perturbation measurement system. In the first experiment, it was found that 40 secs of oral reading provided relatively stable long-term speaker-characterizing parameters of frequency perturbation for the phonations of healthy speakers. In the second experiment, it was demonstrated that low-frequency phase distortions have a substantial adverse affect on the measurement of perturbations in time domain speech waveforms and should be compensated by appropriate techniques. The other experiments presented in Chapter 6 were a series of statistical evaluations in which the perturbation measurement system was applied to the task of differentiating between a control group of speakers thought to be healthy and a group of speakers diagnosed for a variety of laryngeal pathologies. The groups of control and pathological speakers were evaluated by goodness-of-fit statistical procedures and it was found that the sample distributions for most of the parameters were representative of statistically normal distributions. The second statistical evaluation used analysis of variance to determine if the factors of voice condition and sex of speaker had significant effects on the discrimination between control and pathological speakers by intonation and perturbation parameters. It was revealed that the best differentiation between control and pathological groups of speakers is achieved when gender of speaker is also used as a grouping factor. Pearson correlational analysis was completed for each speaker group to determine the types and degrees of correlations amongst the various intonation and perturbation parameters. High, positive correlations were revealed for each type of parameter (i.e. F0 intonation, F0 perturbation and A0

perturbation) which suggests that many of the parameters represent redundant information. In the final set of statistical tests, the Maximum Likelihood principle was used to classify speakers as pathological or control based on multidimensional functions derived from the intonation and perturbation parameters. The results of the classification tasks demonstrated high rates of successful classification of the speakers contained in the test groups of control and pathological speakers. It was concluded that the intonation and perturbation parameters as extracted by the perturbation measurement system are useful for differentiating between groups of healthy speakers and speakers with known pathological conditions of the larynx.

Medium-term Developmental Research

A number of issues are still outstanding in regard to the development and performance of the perturbation measurement system. Firstly, the success of the system for quantifying waveform perturbations is very dependent on the performance of the parallel processing pitch detection algorithm. It would appear from the initial results that the parallel processor produces useful data for measuring F0 and A0 perturbation parameters which can adequately discriminate between healthy and pathological speakers. Some preliminary investigations into the performance of the parallel processor have been presented in Sections 3.2 and 3.3 above as well as in Laver et al. (1982) and Hiller et al. (1983). In these investigations, evaluation of the pitch detector's performance accuracy and reliability were based on comparisons with manual (i.e. visual) pitch extraction and the outputs of other PDAs. More

extensive evaluations of the accuracy and reliability of the parallel processor should be completed, in particular, for irregular signals as produced by pathological speakers. It is recognized that the calibration and assessment of PDA performance is one of the more difficult areas of research in the speech signal processing field. Further research into the performance of the parallel processor as applied to standardized synthetic speech signals should shed some light in this area. The performance of the parallel processor should also be investigated for various noise conditions, in particular, those conditions associated with the clinical environments where the system has its greatest potential use. Secondly, further research is required into the measurement of perturbation parameters derived from the F0 and A0 contours produced by the parallel processor. Since the concept of trend line analysis of perturbations seems appropriate for the evaluation of connected samples of speech, other long-term statistics (e.g. the median measure of central tendency) could provide additional information on perturbatory activity about the trend line. The usefulness of the Directional Perturbation Factors, as demonstrated in the pattern recognition classification tasks, would also suggest the need for further investigation of cycle-to-cycle perturbation analysis of contours extracted from samples of connected speech. Thirdly, there is the issue of the general form of the system -- should it remain as a suite of computer programs or would the system benefit from a hardware implementation? In general, a substantial decrease in analysis time would be realized if the perturbation measurement system were, in some part, converted to hardware form. Two immediate benefits of hardware implementation of the system include a substantial increase in sampling resolution and the elimination of

the need to compensate for recorder-induced phase distortions of the input signal. The development of the perturbation measurement system as a hardware device would enable its direct use in clinical settings.

Future Applications of the Perturbation Measurement System

The initial success achieved by the development and evaluation of the measurement system warrants further investigation into the acoustic analysis of waveform perturbations. Three potential applications of the system for quantifying laryngeal function associated with healthy and pathological phonation include:

1. Screening — The perturbation measurement system is applied as a screening tool to an unselected population of speakers in order to separate those individuals with possible laryngeal pathologies from those who probably do not have disorders. The development of the system for voice screening requires research in a number of areas. Firstly, there is the practical task of enlarging the sample sizes of the control and pathological groups (presented in Chapter 6 above), thus increasing the power of the statistical tests used for multidimensional classification of the speakers. Secondly, clinical trials are required in which new unknown cases are classified for condition of voice based on the multidimensional functions developed for the known cases of laryngeal state. In the clinical trials, emphasis should be placed on the detection of laryngeal pathology in the early stages of development which could enable prompt treatment of the disorder. Thirdly, the tuning of the screening procedures

towards high rates of correct detection of laryngeal pathology should be studied. The overlaps between the groups of healthy and pathological speakers for all of the acoustic parameters indicate that a bias towards, for example, a 100% rate of correct detection will produce increased numbers of false alarms as a consequence. Therefore, the clinical application of the system for screening should be somewhat limited in scope, for instance, as part of existing programs for screening in schools, hospitals, "well-man/well-woman" clinics, etc., where higher false alarm rates would be considered acceptable.

2. Diagnostic Support -- In this application, the perturbation measurement system could be one part of a battery of tests (acoustic and otherwise) used in the differential diagnosis of suspected cases of laryngeal pathology. This task takes the classification procedures used for screening one step further in that a variety of laryngeal pathologies must be discriminated by the acoustic parameters. Some preliminary observation of the data collected thus far suggests that the intonation and perturbation parameters tend to cluster (in the multidimensional sense) into patterns particular to a given structural class of laryngeal pathology (e.g. lesions of the epithelial versus lamina propria tissues of the vocal folds). To investigate this clustering behavior further, more data should be collected for each type of laryngeal pathology in order that classification techniques such as the Maximum Likelihood principle can be properly applied to the discrimination of pathologies.

3. Longitudinal Monitoring -- The perturbation measurement system is

used for the continuing assessment of laryngeal function displayed by diagnosed cases of pathology. In this application, single case studies are used to quantify the progressive nature of laryngeal pathology. In the first instance, the acoustic parameters are measured on repeated occasions in order to monitor therapeutic progress following medical treatment (e.g. surgery, chemotherapy or radiotherapy) or during voice therapy. The progressive deterioration of laryngeal function may also be monitored by continuing assessment of perturbation parameters. Both the diagnostic support and longitudinal monitoring applications suggest the notion of an acoustic vocal profile (see, for example, Davis 1976) which characterizes each pathology and provides a useful record for assessment.

The positive results found for the perturbation measurement system in these early investigations indicate that this acoustic analysis technique has a potentially valuable contribution to offer in the assessment of laryngeal function.

REFERENCES

R E F E R E N C E S

ICASSP = Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing

IEEE Trans. ASSP = IEEE Transactions on Acoustics, Speech and Signal Processing

IEEE Trans. AU = IEEE Transactions on Audio and Electroacoustics

JASA = Journal of the Acoustical Society of America

JSHR = Journal of Speech and Hearing Research

Proc. IEEE = Proceedings of the IEEE

STL-QPSR = Speech Transmission Laboratory, Quarterly Progress and Status Report

Anderson F. (1960); An experimental pitch indicator for training deaf scholars. JASA, 32, 1065-1074.

Askenfelt A. and Hammarberg B. (1980); Speech waveform perturbation analysis. STL-QPSR, 4, 40-49.

Askenfelt A. and Hammarberg B. (1981); Speech waveform perturbation analysis revisited. STL-QPSR, 4, 49-68.

Askenfelt A. and Sjölin A. (1980); Voice analysis in depressed patients: rate of change of fundamental frequency related to mental state. STL-QPSR, 2-3, 71-84.

Atal B. (1976); Automatic recognition of speakers from their voices. Proc. IEEE, 64, 460-475.

Benjamin B.J. (1981); Frequency variability in the aged voice. J. Gerontology, 36, 722-726.

Berouti M., Childers D.G. and Paige A. (1977); Correction of tape recorder distortion. ICASSP-77, 397-400.

Bogert B.P., Healy M.J.R. and Tukey J.W. (1963); The quefrency alanysis of time series echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe-cracking. In Rosenblatt M. (Ed.) Proceedings of the Symposium on Time Series Analysis, New York:J. Wiley and Sons, 209-243.

Bryce D. (1974); Differential Diagnosis and Treatment of Hoarseness, Springfield:Charles Thomas Pub.

Coleman R.P. (1969); Effect of median frequency levels upon the roughness of jittered stimuli. JSHR, 12, 330-336.

Davis S.B. (1976); Computer evaluation of laryngeal pathology based on inverse filtering of speech. Speech Communication Research Lab, Santa Barbara CA, SCRL Monograph 13.

Davis S.B. (1979); Acoustic characteristics of normal and pathological voices. In Lass N.J. (Ed.), Speech and Language: Advances in Basic Research and Practice, New York:Academic Press, 1, 273-338.

Deal R. and Emanuel F. (1978); Some waveform and spectral features of vowel roughness. JSHR, 21, 250-264.

Dolansky L.O. (1955); An instantaneous pitch-period indicator. JASA, 27, 67-72.

Dubnowski J.J., Schafer R.W. and Rabiner L.R. (1976); Real-time digital hardware pitch detector. IEEE Trans. ASSP-24, 2-8.

Duda R.O. and Hart P.E. (1973); Pattern Classification and Scene Analysis. New York:John Wiley and Sons.

Duifhuis H., Willems L.F. and Sluyter R.J. (1982); Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception. JASA, 71, 1568-1580.

Emanuel F.W. and Sansone Jr. F.E. (1969); Some spectral features of 'normal' and simulated 'rough' vowels. Folia Phoniatrica, 21, 401-415.

Fairbanks G. (1960); Voice and Articulation Drillbook. New York:Harper Brothers.

Fant G. (1960); Acoustic Theory of Speech Production. 's-Gravenhage:Mouton and Co.

Fant G. (1979a); Voice source analysis - a progress report. STL-QPSR, 3-4, 31-53.

Fant G. (1979b); Glottal source and excitation analysis. STL-QPSR, 1, 85-107.

Gold B. (1962); Computer program for pitch extraction. JASA, 34, 916-921.

Gold B. (1964); Note on buzz-hiss detection. JASA, 36, 1659-1661.

Gold B. and Rabiner L.R. (1969); Parallel processing techniques for estimating pitch periods of speech in the time domain. JASA, 46, 442-448.

Goldstein J.L. (1973); An optimum processor theory for the central formation of the pitch of complex tones. JASA, 54, 1496-1516.

Green N. (1972); Automatic speaker recognition using pitch measurements in conversational speech. JSRU Report No. 1000, Joint Speech Research Unit, Ruislip, Middlesex.

Hanson R.J. (1978); A two-state model of F0 control. JASA, 64, 543-544.

Harrington J. and Hiller S. (1984); Laryngeal tension and

stuttering. Edinburgh University Department of Linguistics, Work in Progress, 17, 127-134.

Hecker M. and Kreul E. (1971); Descriptions of the speech of patients with cancer of the vocal folds. Part 1 : measures of fundamental frequency. JASA, 49, 1275-1282.

Hess W. (1980); Pitch determination of speech signals - a survey. In Simon J.C. (Ed.) Spoken Language Generation and Understanding. Proceedings of the NATO Advanced Study Institute, Bonas, France 1979.

Hess W. (1982); Algorithms and devices for pitch determination of speech signals. Phonetica, 39, 219-240.

Hess W. (1983); Pitch Determination of Speech Signals: Algorithms and Devices. Berlin:Springer-Verlag.

Hiller S.M., Laver J. and Mackenzie J. (1983); Automatic analysis of waveform perturbations in connected speech. Edinburgh University Department of Linguistics, Work in Progress, 16, 40-69.

Hiller S.M., Laver J. and Mackenzie J. (1984); Durational Aspects of long-term measurements of fundamental frequency perturbations in connected speech. Edinburgh University Department of Linguistics, Work in Progress, 17, 59-76.

Hollien H., Michel J., and Doherty E.T. (1973); A method for analyzing vowel jitter in sustained phonation. J. Phonetics, 1, 85-91.

Holmes J.N. (1973); The influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. IEEE Trans. AU-21, 298-305.

Holmes J.N. (1975); Low-frequency phase distortion of speech recordings. JASA, 58, 747-749.

Horii Y. (1975); Some statistical characteristics of voice fundamental frequency. JSHR, 18, 192-201.

Horii Y. (1979); Fundamental frequency perturbation observed in sustained phonation. JSHR, 22, 5-19.

Horii Y. (1980); Vocal shimmer in sustained phonation. JSHR, 23, 202-209.

Horii Y. (1982); Jitter and shimmer differences among sustained vowel phonations. JSHR, 25, 12-14.

Horii Y. (1985); Jitter and shimmer in vocal fry phonation. Folia Phoniatrica, 37, 81-86.

Iwata S. and von Leden H. (1970); Pitch perturbations in normal and pathological voices. Folia Phoniatrica, 22, 413-424.

Kanal L. (1974); Patterns in pattern recognition: 1968-1974. IEEE Transactions on Information Technology, IT-20, 697-722.

Kane M. and Wellen C.J. (1985); Acoustical measurements and clinical judgements of vocal quality in children with vocal nodules. Folia Phoniatrica, 37, 53-57.

Kasprzyk P.L. and Gilbert H.R. (1975); Vowel perturbation as a function of vowel height. JASA, 57, 1545-1546.

Kasuya H., Kobayashi Y. and Kobayashi T. (1983); Characteristics of pitch period and amplitude perturbations in pathologic voice. ICASSP-83, 1372-1375.

Kempster G.B. and Kistler D.J. (1983); Selected acoustic characteristics of pathological and normal speakers: a reanalysis. JSHR, 26, 159-160.

Kitajima K. and Gould W.J. (1976); Vocal shimmer in sustained phonations of normal and pathological voices. Annals of Otolaryngology, 85, 377-381.

Kitajima K., Tanabe M. and Isshiki N. (1975); Pitch perturbations in normal and pathological voice. Studia Phonetica, 9, 25-32.

Koike Y. (1973); Application of some acoustic measures for the evaluation of laryngeal dysfunction. Studia Phonetica, 7, 17-23.

Koike Y., Takahashi H. and Calcaterra T.C. (1977); Acoustic measures for detecting laryngeal pathology. Acta Otolaryngology, 84, 105-117.

Kubzdela, H. (1976); An analogue fundamental frequency extractor. In Jassem W. (Ed.) Speech Analysis and Synthesis, Warsaw: Polish Academy of Sciences, 4, 269-279.

Laver J. (1980); The Phonetic Description of Voice Quality. Cambridge:Cambridge University Press.

Laver J. and Hanson R.J. (1981); Describing the normal voice. In Darby J. (Ed.) Speech Evaluation in Psychiatry, New York:Grune and Stratton, 51-78.

Laver J., Hiller S. and Hanson R. (1982); Comparative performance of pitch detection algorithms on dysphonic voices. ICASSP-82, 192-195.

Laver, J., Hiller S. and Mackenzie J. (1984); Acoustic analysis of vocal fold pathology. Proceedings of The Institute of Acoustics, 6, 235-452.

Laver J., Wirz S., Mackenzie J., and Hiller S. (1981); A perceptual protocol for the analysis of vocal profiles. Edinburgh University Dept. of Linguistics, Work in Progress, 14, 139-155.

von Leden H. and Koike Y. (1970); Detection of laryngeal disease by computer techniques. Archives of Otolaryngology, 91, 33-10.

Lieberman P. (1961); Perturbations in vocal pitch. JASA, 33, 597-603.

Lieberman P. (1963); Some acoustic measures of the fundamental periodicity of normal and pathological larynges. JASA, 35, 344-353.

Ludlow C.L., Coulter D.C. and Gentges F.H. (1983a); The differential sensitivity of frequency perturbation to laryngeal neoplasms and neuropathologies. In Bless D.M. and Abbs J.H. (Eds.) Vocal Fold Physiology: contemporary research and clinical issues. San Diego: College Hill, 381-392.

Ludlow C.L., Coulter D.C. and Gentges F.H. (1983b); The effects of change in vocal fold morphology on phonation. In Lawrence V.L. (Ed.) Transcripts of the Eleventh Symposium Care of the Professional Voice, Part I: Scientific Sessions: Papers. The Voice Foundation, New York, NY., 77-89.

Ludlow C.L., Naunton R.F. and Bassich C.J. (1984); Procedures for the selection of spastic dysphonia patients for recurrent laryngeal nerve section. Otolaryngology Head and Neck Surgery, 92, 24-31.

Luksaneeyanawin S. (1984); The tonal behavior of one-word utterances: the interplay between tone and intonation in Thai. Edinburgh University Department of Linguistics, Work in Progress, 17, 16-30.

Mackenzie J., Laver J. and Hiller S.M. (1983); Structural pathologies of the vocal folds and phonation. Edinburgh University Department of Linguistics, Work in Progress, 16, 80-117.

Makhoul J. (1975); Linear prediction: a tutorial review. Proc. IEEE, 63, 561-580.

Markel J.D. (1972); The SIFT algorithm for fundamental frequency estimation. IEEE Trans. AU-20, 367-377.

Markel J.D. and Davis S.B. (1979); Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. IEEE Trans. ASSP-27, 74-82.

Markel J.D. and Gray A.H. (1976); Linear Prediction of Speech, Berlin:Springer-Verlag.

Markel J., Oshika B. and Gray G. (1977); Long-term feature averaging for speaker recognition. IEEE Trans. ASSP-25, 330-337.

Martin P. (1981); Détection de F0 par intercorrélation avec une fonction peigne. Journées d'Etude sur la Parole 12, 221-232.

Martin P. (1982); Comparison of pitch detection by cepstrum and spectral comb analysis. ICASSP-82, 180-183.

McClellan J. (1975); FIR design program. In Rabiner L.R. and Gold B. (Eds.) Theory and Application of Digital Signal Processing, New Jersey:Prentice-Hall, 194-204.

McDonald W.E., Zyski B.J., Johns M.E. and Bull G.L. (1981); Adjunctive use of perturbation analysis for objective assessment of laryngeal surgery. IEEE Proceedings of the 5th Annual Symposium on Computer Applications in Medical Care.

McGonegal C.A., Rabiner L.R. and Rosenberg A.E. (1977); A subjective evaluation of pitch detection methods using LPC synthesized speech. IEEE Trans. ASSP-25, 221-229.

McKinney N.P. (1965); Laryngeal frequency analysis for linguistic research. Communication Sciences Laboratory, University of Michigan, Ann Arbor, Research Report 14.

Mead K.O. (1974); Identification of speakers from fundamental frequency contours in conversational speech. JSRU Report No. 1002, Joint Speech Research Unit, Ruislip, Middlesex.

Miller N.J. (1975); Pitch detection by data reduction. IEEE Trans. ASSP-23, 72-79.

Moorer J.A. (1974); The optimum comb method of pitch period analysis of continuous digitized speech. IEEE Trans. ASSP-22, 330-338.

Murray T. and Doherty E.T. (1980); Selected acoustic characteristics of pathological and normal speakers. JSHR, 23, 361-369.

Murray T. and Doherty E.T. (1983); Selected acoustic characteristics of pathological and normal speakers: a reanalysis: a reply to Kempster and Kistler. JSHR, 26, 160.

Nagy G. (1968); Classification algorithms for pattern recognition. IEEE Trans. AU-16, 203-212.

Nolan N. (1983); The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press.

Noll A.M. (1967); Cepstrum pitch determination. JASA, 41, 293-309.

Noll A.M. (1970); Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In Microwave Institute (Ed.) Symposium on Computer Processing in Communication, New York:University of Brooklyn Press, 19, 779-797.

Olsen G.H. (1982); Modern Electronics Made Simple. London:Heinemann.

Paliwal K.K. and Rao P.V.S. (1983); A synthesis-based method for pitch extraction. Speech Communication, 2.

Perkins W.H. (1977); Vocal function: a behavioral analysis. In Travis L.E. (Ed.) Handbook of Speech Pathology and Audiology, New York:Appleton Croft, 481-503.

Rabiner L.R. (1977); On the use of autocorrelation analysis for pitch detection. IEEE Trans. ASSP-25, 24-33.

Rabiner L.R., Cheng M.J., Rosenberg A.E., and McGonegal C.A. (1976); A comparative study of several pitch detection algorithms. IEEE Trans. ASSP-24, 399-413.

Rabiner L.R. and Sambur M.R. (1977); Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. IEEE Trans. ASSP-25, 338-343.

Rabiner L.R., Sambur M.R. and Schmidt C.E. (1975); Applications of nonlinear smoothing algorithm to speech processing. IEEE Trans. ASSP-23, 552-557.

Rabiner L.R. and Schafer R.W. (1978); Digital Processing of Speech Signals. New Jersey:Prentice Hall.

Rabiner, L.R., Schmidt, C.E. and Atal, B.S. (1977); Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone-quality speech. Bell System Technical J., 56, 455-482.

Ramig L.A. and Ringel R.L. (1983); Effects of physiological aging on selected acoustic characteristics of voice. JSHR, 26, 22-30.

Reddy D.R. (1967); Pitch period determination of speech sounds. Communications of the Association of Computing Machines, 10, 343-348.

Rosenberg A.E. (1976); Automatic speaker verification - a review. Proc. IEEE, 64, 475-487.

Ross M.J., Shaffer H.L., Cohen A., Freudberg R. and Manley H.J. (1974); Average magnitude difference function pitch extractor. IEEE Trans. ASSP-22, 353-361.

Sambur M.R. (1978); Adaptive noise canceling for speech signals. IEEE Trans. ASSP-26, 419-423.

Sambur M.R. and Rabiner L.R. (1975); A speaker independent digit-recognition system. Bell System Technical J., 54, 81-102.

Sansone F.E. and Emanuel F.W. (1970); Spectral noise levels and roughness ratings for normal and simulated rough vowels produced by adult males. JSHR, 13, 489-502.

Schroeder M.R. (1968); Period histogram and product spectrum: new

methods for fundamental-frequency measurement. JASA, 43, 829-834.

Schroeder M.R. and Atal B.S. (1962); Generalized short-time power spectra and autocorrelation functions. JASA, 34, 1679-1683.

Seneff S. (1978); Real time harmonic pitch detector. IEEE Trans. ASSP-26, 358-364.

Sherman D. and Linke E. (1952); The influence of certain vowel types on degree of harsh vowel quality. J. Speech Disorders, 17, 401-408.

Siegel L.J. and Bessey A.C. (1982); Voiced/unvoiced/mixed excitation classification of speech. IEEE Trans. ASSP-30, 451-460, 1982.

Siegel L.J. and Steiglitz K.S. (1976); A pattern classification algorithm for the voiced/unvoiced decision. ICASSP-76, 326-329.

Smith B.E., Weinberg B., Lawrence L.F., and Horii Y. (1978); Vocal roughness and jitter characteristics of vowels produced by esophageal speakers. JSHR, 21, 240-249.

Smith C.P. (1954); Device for extracting the excitation function from speech signals. United States Patent No. 2,691,137. Issued Oct.5, 1954; filed June 27, 1952; reissued 1956.

Smith C.P. (1957); Speech data reduction: voice communications by means of binary signals at rates under 1000 bits/sec. AFCRC, Bedford MA; DDC-AD-117290.

Smith W.R. and Lieberman P. (1969); Computer diagnosis of laryngeal lesion. Computers and Biomedical Research, 2, 291-303.

Sondhi M.M. (1968); New methods for pitch extraction. IEEE Trans. AU-16, 262-266.

Sorenson D., Horii Y. and Leonard R. (1980); Effects of laryngeal topical anesthesia on voice fundamental frequency perturbation. JSHR, 23, 274-283.

Sreenivas T.V. and Rao P.V.S. (1979); Pitch extraction from corrupted harmonics of the power spectrum. JASA, 65, 223-228.

Steffen-Batog M., Jassem W. and Gruszka-Koscielak, H. (1970); Statistical distribution of short-term F0 values as a personal voice characteristic. In Jassem W. (Ed.) Speech Analysis and Synthesis, Warsaw: Polish Academy of Sciences, 2, 195-206.

Terhardt E. (1974); Pitch, consonance, and harmony. JASA, 55, 1061-1069.

Terhardt E., Stoll G. and Seewann M. (1982a); Pitch of complex signals according to virtual-pitch theory: tests, examples,

and predictions. JASA, 71, 671-678.

Terhardt E., Stoll G. and Seewann M. (1982b); Algorithm for extraction of pitch and pitch salience from complex tonal signals. JASA, 71, 679-688.

Tucker W.H. and Bates R.H.T. (1978); A pitch estimation algorithm for speech and music. IEEE Trans. ASSP-26, 597-604.

Un C.K. and Yang S.C. (1977); A pitch extraction algorithm based on LPC inverse filtering and AMDF. IEEE Trans. ASSP-25, 565-572.

Ungeheuer G. (1963); Zur Periodizitätsanalyse phonetischer Signale (Gegenwärtiger Stand der Entwicklung von Tonhöhenschreibern). Phonetica, 10, 174-186.

von Leden H. and Koike Y. (1970); Detection of laryngeal disease by computer techniques. Archives of Otolaryngology, 91, 33-10.

Wakita H. (1976); Instrumentation for the study of speech acoustics. In Lass N.J. (Ed.) Comtemporary Issues in Experimental Phonetics, New York:Academic Press Inc., 3-40.

Wendahl R.W. (1963); Laryngeal synthesis of harsh voice quality. Folia Phoniatrica, 15, 241-250.

Wendahl R.W. (1966a); Some parameters of auditory roughness. Folia Phoniatrica, 18, 26-32.

Wendahl R.W. (1966b); Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness. Folia Phoniatrica, 18, 98-108.

Wilcox K.A. and Horii Y. (1980); Age and changes in vocal jitter. J. of Gerontology, 35, 194-198.

Yaggi L.A. (1962); Full duplex digital vocoder. Scientific report No. 1. Texas Instruments, Dallas TX; SP14-A62; DDC-AD-282986.

Yaggi L.A. (1963); Full duplex digital vocoder, final report No. 2. Texas Instruments, Dallas TX; Rept. No. SP16-A63.

Zyski B.J., Bull G.L., McDonald W.E. and Johns M.E. (1984); Perturbation analysis of normal and pathologic larynges. Folia Phoniatrica, 36, 190-198.

A P P E N D I C E S

APPENDIX 1

ANOVA tables for the 10 acoustic parameters; factors include VOICE CONDITION (VCOND) and SEX.

a) FO-AV

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	326553.998	2	163276.999	382.887	0.000
VCOND	135.165	1	135.165	0.317	0.574
SEX	326508.238	1	326508.238	765.668	0.000
2-WAY INTERACTION	7143.388	1	7143.388	16.751	0.000
EXPLAINED	333697.386	3	111232.462	260.842	0.000
RESIDUAL	96374.537	226	426.436		
TOTAL	430071.923	229	1878.043		

b) FO-DEV

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	18373.679	2	9186.839	148.856	0.000
VCOND	225.337	1	225.337	3.651	0.057
SEX	17999.705	1	17999.705	291.653	0.000
2-WAY INTERACTION	54.622	1	54.622	0.885	0.348
EXPLAINED	18428.301	3	6142.767	99.533	0.000
RESIDUAL	13947.855	226	61.716		
TOTAL	32376.156	229	141.381		

APPENDIX 1

ANOVA tables for the 10 acoustic parameters (continued); factors include VOICE CONDITION (VCOND) and SEX.

c) J-DEVEX

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	127.121	2	63.560	4.857	0.009
VCOND	38.016	1	38.016	2.905	0.090
SEX	92.796	1	92.796	7.091	0.008
2-WAY INTERACTION	11.247	1	11.247	0.859	0.355
EXPLAINED	138.367	3	46.122	3.525	0.016
RESIDUAL	2957.439	226	13.086		
TOTAL	3095.806	229	13.519		

d) J-AVEX

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	56.358	2	28.179	8.423	0.000
VCOND	45.399	1	45.399	13.570	0.000
SEX	12.430	1	12.430	3.715	0.055
2-WAY INTERACTION	4.820	1	4.820	1.441	0.231
EXPLAINED	61.178	3	20.393	6.095	0.001
RESIDUAL	756.095	226	3.346		
TOTAL	817.273	229	3.569		

APPENDIX 1

ANOVA tables for the 10 acoustic parameters (continued); factors include VOICE CONDITION (VCOND) and SEX.

e) J-RATEX

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	2009.374	2	1004.687	18.457	0.000
VCOND	1194.472	1	1194.472	21.944	0.000
SEX	878.758	1	878.758	16.144	0.000
2-WAY INTERACTION	94.156	1	94.156	1.730	0.190
EXPLAINED	2103.530	3	701.177	12.881	0.000
RESIDUAL	12302.029	226	54.434		
TOTAL	14405.560	229	62.906		

f) J-DPF

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	2408.035	2	1204.018	42.430	0.000
VCOND	1421.441	1	1421.441	50.092	0.000
SEX	1063.218	1	1063.218	37.468	0.000
2-WAY INTERACTION	76.995	1	76.995	2.713	0.101
EXPLAINED	2485.030	3	828.343	29.191	0.000
RESIDUAL	6413.089	226	28.377		
TOTAL	8898.119	229	38.856		

APPENDIX 1

ANOVA tables for the 10 acoustic parameters (continued); factors include VOICE CONDITION (VCOND) and SEX.

g) S-DEVEX

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	14836.974	2	7418.487	1.157	0.316
VCOND	3217.814	1	3217.814	0.502	0.479
SEX	11217.147	1	11217.147	1.750	0.187
2-WAY INTERACTION	4169.539	1	4169.539	0.650	0.421
EXPLAINED	19006.513	3	6335.504	0.988	0.399
RESIDUAL	1448892.082	226	6411.027		
TOTAL	1467898.595	229	6410.038		

h) S-AVEX

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	1337.563	2	668.781	23.232	0.000
VCOND	1013.758	1	1013.758	35.216	0.000
SEX	361.373	1	361.373	12.553	0.000
2-WAY INTERACTION	5.434	1	5.434	0.189	0.664
EXPLAINED	1342.997	3	447.666	15.551	0.000
RESIDUAL	6505.859	226	28.787		
TOTAL	7848.856	229	34.274		

APPENDIX 1

ANOVA tables for the 10 acoustic parameters (continued); factors include VOICE CONDITION (VCOND) and SEX.

i) S-RATEX

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	14089.512	2	7044.756	87.342	0.000
VCOND	8245.946	1	8245.946	102.235	0.000
SEX	6292.563	1	6292.563	78.016	0.000
2-WAY INTERACTION	67.317	1	67.317	0.835	0.362
EXPLAINED	14156.829	3	4718.943	58.506	0.000
RESIDUAL	18228.476	226	80.657		
TOTAL	32385.305	229	141.421		

j) S-DPF

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	10476.908	2	5238.454	132.106	0.000
VCOND	8653.176	1	8653.176	218.221	0.000
SEX	2086.332	1	2086.332	52.614	0.000
2-WAY INTERACTION	51.028	1	51.028	1.287	0.258
EXPLAINED	10527.935	3	3509.312	88.500	0.000
RESIDUAL	8961.653	226	39.653		
TOTAL	19489.588	229	85.107		

APPENDIX 2 PEARSON CORRELATION COEFFICIENTS: MALE CONTROL GROUP

	FOAV	FODEV	JDEVEX	J-AVEX	JRATEX	JDPF	SDEVEX	SAVEX	SRATEX	SDPF
FOAV	1.00	0.75 ****	-0.06	0.04	-0.10	-0.21	-0.01	-0.34 ****	-0.45 ****	-0.46 ****
FODEV		1.00	0.17	0.25	0.14	-0.08	-0.10	-0.10	-0.18	-0.30
JDEVEX			1.00	0.89 ****	0.49 ****	0.19	0.17	0.52 ****	0.28	0.02
JAVEX				1.00	0.78 ****	0.50 ****	0.15	0.61 ****	0.37 ****	0.14
JRATEX					1.00	0.89 ****	0.14	0.70 ****	0.64 ****	0.46 ****
JDPF						1.00	0.06	0.56 ****	0.71 ****	0.63 ****
SDEVEX							1.00	0.57 ****	-0.03	-0.05
SAVEX								1.00	0.61 ****	0.44 ****
SRATEX									1.00	0.80 ****
SDPF										1.00

**** -- Significant Correlation

APPENDIX 2 PEARSON CORRELATION COEFFICIENTS: FEMALE CONTROL GROUP

	FOAV	FODEV	JDEVEX	JAVEX	JRATEX	JDPF	SDEVEX	SAVEX	SRATEX	SDPF
FOAV	1.00	0.61 ****	-0.29	-0.38 ****	-0.50 ****	-0.59 ****	0.01	-0.34 ****	-0.49 ****	-0.50 ****
FODEV		1.00	0.32	0.23	0.08	-0.12	0.23	0.23	0.01	-0.00
JDEVEX			1.00	0.96 ****	0.84 ****	0.64 ****	0.22	0.80 ****	0.64 ****	0.63 ****
JAVEX				1.00	0.94 ****	0.77 ****	0.14	0.80 ****	0.74 ****	0.74 ****
JRATEX					1.00	0.90 ****	0.02	0.73 ****	0.82 ****	0.78 ****
JDPF						1.00	-0.10	0.57 ****	0.80 ****	0.83 ****
SDEVEX							1.00	0.61 ****	0.03	0.05
SAVEX								1.00	0.68 ****	0.64 ****
SRATEX									1.00	0.89 ****
SDPF										1.00

**** -- Significant Correlation

APPENDIX 2 PEARSON CORRELATION COEFFICIENTS: MALE PATHOLOGICAL GROUP

	FOAV	FODEV	JDEVEX	JAVEX	JRATEX	JDPF	SDEVEX	SAVEX	SRATEX	SDPF
FOAV	1.00	0.75 ****	-0.12	0.02	-0.05	-0.15	-0.22	-0.33	-0.48 ****	-0.51 ****
FODEV		1.00	0.39 ****	0.47 ****	0.37 ****	0.21	0.03	0.13	-0.01	-0.15
JDEVEX			1.00	0.92 ****	0.79 ****	0.69 ****	0.46 ****	0.71 ****	0.61 ****	0.46 ****
JAVEX				1.00	0.93 ****	0.82 ****	0.37 ****	0.69 ****	0.58 ****	0.35
JRATEX					1.00	0.95 ****	0.30	0.65 ****	0.59 ****	0.33
JDPF						1.00	0.32	0.65 ****	0.60 ****	0.41 ****
SDEVEX							1.00	0.73 ****	0.28	0.29
SAVEX								1.00	0.75 ****	0.61 ****
SRATEX									1.00	0.86 ****
SDPF										1.00

**** -- Significant Correlation

APPENDIX 2 PEARSON CORRELATION COEFFICIENTS: FEMALE PATHOLOGICAL GROUP

	FOAV	FODEV	JDEVEX	JAVEX	JRATEX	JDPF	SDEVEX	SAVEX	SRATEX	SDPF
FOAV	1.00	0.64 ****	-0.30	-0.32	-0.43 ****	-0.44 ****	-0.03	-0.44 ****	-0.42 ****	-0.49 ****
FODEV		1.00	0.33	0.27	0.13	-0.01	0.04	0.03	-0.10	-0.25
JDEVEX			1.00	0.97 ****	0.89 ****	0.78 ****	0.12	0.82 ****	0.67 ****	0.57 ****
JAVEX				1.00	0.96 ****	0.86 ****	0.06	0.86 ****	0.73 ****	0.63 ****
JRATEX					1.00	0.95 ****	0.05	0.88 ****	0.74 ****	0.65 ****
JDPF						1.00	0.01	0.81 ****	0.79 ****	0.74 ****
SDEVEX							1.00	0.39 ****	-0.13	-0.15
SAVEX								1.00	0.69 ****	0.62 ****
SRATEX									1.00	0.94 ****
SDPF										1.00

**** -- Significant Correlation

APPENDIX 3

COMPARATIVE PERFORMANCE OF PITCH DETECTION ALGORITHMS ON DYSPHONIC VOICES

John Laver — Phonetics Laboratory, University of Edinburgh, Scotland

Steven Hiller — Phonetics Laboratory, University of Edinburgh, Scotland

Robert Hanson — Bell Laboratories, Murray Hill, New Jersey, U.S.A.

ABSTRACT

Current pitch detection algorithms run into difficulties when used on dysphonic voices. Two major sources of difficulty are the presence in the phonatory output of frictional, non-harmonic energy (in whispery voices), and microperturbatory fundamental frequency jitter and amplitude shimmer (in harsh and creaky voices). For adequate performance on dysphonic voices, pitch detection algorithms should have the following characteristics:

1. work on acoustic recordings from men, women and children
2. be noise resistant
3. work on continuous speech.

Measures of pitch perturbation are defined. Three pitch detection algorithms were applied to the speech of dysphonic speakers as well as a control group of speakers. Two detectors work in the time domain (simplified inverse filter tracking (1) and a parallel processing method (2)), and one in the frequency domain (cepstral pitch detection (3)). Their comparative performance on perceptually rated clinical material is discussed.

INTRODUCTION

Automatic acoustic analysis of clinical recordings, helpful in the description, diagnosis and rehabilitation of voice and speech disorders, is a rapidly growing research area. Algorithms designed to work on the acoustic characteristics of normal speech do not always work very effectively on the perturbed acoustic signals of dysphonic speech. This is particularly true of the pitch detection algorithms available to clinical research.

The ideal acoustic material for effective pitch detection would be a recording with a good signal-to-noise ratio of a young adult male with optimally efficient laryngeal vibration characteristics. Acoustically and physiologically, efficient phonation of this sort can be described as phonation where the vibration of the true vocal folds is regularly periodic, efficient in air use, without audible friction, with the folds in full glottal vibration under moderate values on all myodynamic parameters. In this type of vibration, the larynx pulse shape is approximately triangular, with maximum excitation of the supralaryngeal vocal tract occurring during the closing phase of the glottal cycle, and with the closing phase lasting for about 33% of the cycle. The spectral slope of the glottal waveform is -10 dB below 250 Hz and -12 dB above it. The larynx pulse has a very limited range of jitter and shimmer, and these microperturbations of fundamental frequency and amplitude have a normal distribution, with a standard deviation 2% or less of the mean frequency and amplitude. The range of fundamental frequency is from 50 to 250 Hz (4).

These are ideal characteristics. In dysphonic voices many, and sometimes all of the above factors are distorted, with consequent problems for automatic pitch detection. The two major sources of difficulty are the presence in the laryngeal waveform of fricative, non-harmonic energy, and of frequent and substantial perturbations of fundamental frequency and amplitude, giving excessive jitter and shimmer. In perceptual terms, the addition of fricative energy to the laryngeal waveform gives a "whispery" effect to the phonatory quality; the addition of excessive jitter and shimmer gives a "harsh" effect; and when the jitter and shimmer create a pulse-grouping tendency on the laryngeal waveform, characteristically with a short-long alternation of period durations, a "creaky" effect (also called "vocal fry") is produced. Whispery, harsh and creaky voices are extremely frequent in speech therapy clinics. It is also the case that the majority of speakers in the general population have voices which show phonatory inefficiency in varying degrees, particularly with regard to whisperiness and creakiness (4).

A current project in the Phonetics Laboratory at the University of Edinburgh is testing a descriptive scheme for characterising subjects' voices, in terms of perceptual and acoustic profiles, from a wide range of speech disorders. These disorders include not only dysphonia, but also profound hearing loss, cerebral palsy and Down's Syndrome, all of which are associated with inefficient phonation.

Given that the principal focus of the project is clinical research, rather than the development of signal processing techniques, it was necessary to minimise the amount of time spent writing special programs for the acoustic analysis facility mounted on the laboratory PDP 11/40 computer. The programs readily available for automatic pitch detection included one working in the time domain (simplified inverse filter tracking (1), and one in the frequency domain (cepstral pitch determination (3)), implemented in ILS, the Signal Technology

(Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing 1982, 192-195, 1982).

speech signal processing package, as SIF and API respectively. To broaden the choice slightly, a version (PGR) of a parallel processing method working in the time domain (2) was also included.

CRITERIA FOR PITCH DETECTION ALGORITHM CHARACTERISTICS

In setting out criteria for the evaluation of the three algorithms, it is necessary to take account of the fact that two different aspects of fundamental frequency behavior need to be quantified in the construction of the acoustic profile of a speaker's voice: intonation and microperturbation. All the algorithms would be more or less suitable for quantifying the first aspect, recording the range, mean, median, mode and variability of the intonational use of fundamental frequency. The smoothing which in varying degrees is inherent in all three algorithms is beneficial in recovering the intonationally-relevant pitch contour and minimising the short-term microperturbatory deviations from the contour, which are properly regarded as characterising personal phonatory quality and irrelevant to intonation. But it is precisely the characteristic personal quality of a speaker's voice, together with the intonational statistics that were mentioned above, that the project needs to capture as part of the overall acoustic profile. In terms of general criteria for adequate performance on dysphonic material, it was concluded that a pitch detection algorithm should have the following characteristics:

1. Work on acoustic recordings from men, women and children
2. Be noise resistant
   a. be relatively impervious to poor signal-to-noise ratios arising from poor quality clinical recordings
   b. be resistant to the effects of non-harmonic noise from laryngeal friction
   c. be accurate to within 2% in tracking fundamental frequency (FØ)
   d. retain accuracy over wide FØ and jitter and shimmer ranges
3. Work on continuous speech
   a. have an adequate voicing detector
   b. use a moving-average approach to cope with intonational movements of FØ, and to provide a baseline from which to measure microperturbatory excursion

In the light of the above criteria, PGR was the preferred choice of algorithm. The minimised smoothing involved in this algorithm seemed an acceptable compromise between the objective of tracking the intonational aspects and the objective of adequate accuracy in tracking moment-to-moment perturbational aspects.

PARALLEL PROCESSING METHOD OF MEASURING FUNDAMENTAL FREQUENCY

The Gold and Rabiner (2) method was used in the following manner: a stretch of continuous speech was sampled at a 10 KHz sampling rate. Prior to digitization, the speech was low-pass filtered to 400 Hz, to spectrally flatten the signal. The speech was processed in parallel through six simple pitch detectors, each examining a different aspect of periodicity in the signal. The six estimates of the period were then put to a sophisticated jury system, and the estimate with the majority vote was accepted as the official period. A pre-set level of confidence had to be reached by the vote before the speech was considered voiced and assigned a period.

The correlation between the number of pitch periods yielded by the algorithm and the actual number present in the speech signal is constrained by the size of the shift in successive applications along the sample of the data-inspecting window. When the shift and the period are close in value the algorithm works accurately. But intonational or perturbational movement of FØ beyond a certain limit creates problems. In rapidly rising FØ sequences, when the shift becomes larger than the new period, the algorithm effectively down-samples the actual pulse-train; and in rapidly falling sequences of FØ, when the shift becomes substantially smaller than the new period, the detection system exaggerates the number of periods to be reported. Such an imbalance is less important in tracking the smoothed contour of intonation, but for measuring perturbation accurately, especially in severely dysphonic voices with large perturbatory values, the discrepancy needs to be minimised. A partial solution that was adopted was to make two passes through the data, using an initial 100 point shift. A second pass shift-setting for each voice analysed was then based on the median frequency found in the first pass. An appropriate design modification for the purpose of tracking FØ perturbation would be to make the window-size and shift-setting dynamically pitch-adaptive. Data analysis here was limited, as a compromise between intonational and perturbational interests, to the fundamental frequency values between 50 and 250 Hz.

To check the overall accuracy of the PGR program, a comparison was made of the output on a sample of speech 2.76 seconds long with a visual calculation of the same data. Mean fundamental frequencies calculated by program and by eye were 131.9 Hz and 134.73 Hz respectively. A test of replicability was carried out, by digitizing and analysing a 34.4 second passage from a single recording, on five separate occasions. The range of the mean FØ was 1.0 Hz, over the five repetitions.

Statistical measures extracted from the data fall into two groups - raw frequency data, incorporating minimum, maximum, mean, standard deviation, total range, median and mode values, and frequency perturbation values, using measures called AVEX, SDEVEX, RATEX and DPQEX. The perturbation measures are based on a concept of excursion from a five-point moving mean, expressed as a percentage of that mean. AVEX is the mean excursion, and SDEVEX the standard deviation of the range of excursions. RATEX is defined as the percentage of points in the sample where the excursion is equal to or greater than 3. RATEX is adapted from the moving average approach of Koike,

Takahashi & Calcaterra (5), and from Lieberman's (6) "perturbation factor", except that Lieberman's measure merely reflected absolute differences between two adjacent periods equal to or greater than 0.5 msec. The 3 percent threshold on RATEX means that a signal increasing in FØ by just under 3 percent on each adjacent pulse could rise intonationally from 100 Hz to 150 Hz in less than 0.15 seconds, and still score zero on RATEX. RATEX scores above zero therefore mostly reflect (minor) contributions from sharp intonational corners and (major) contributions from moment-to-moment jitter arising from laryngeal inefficiency (and from algorithmic inefficiency where that is a relevant factor).

DPQEX is adapted from the "directional perturbation factor" of Hecker & Kreul (7), and is defined as the percentage of points in the sample where there is a change in the algebraic sign of the difference between adjacent points, with a 3 percent threshold for the magnitude difference.

RESULTS AND DISCUSSION

20 young adult males, 10 with speech disorders and 10 from a normal control group, were recorded. Passages of continuous speech from 24.29 seconds to 70.12 seconds in length were analysed, using PGR. One speech disorder voice and one control voice of comparable median fundamental frequency were also analysed using SIF and API. All voices had previously been perceptually rated on 6-degree scales of "harshness" (H), "whisperiness" (W) and "creakiness" (C) by trained raters. Table I presents selected intonational and perturbational PGR data for the two groups of subjects, together with their perceptual categorisations.

In Table I, the speech disorder group and the control group overlap with respect to perturbation measures and median FØ, even though the means are well separated. In contrast, harshness and creakiness, as components of voice type, have a relatively complementary distribution. Audible harshness was a criterion for inclusion of subjects in the disorder group; no voice in the randomly-chosen control group displays perceived harshness. Creakiness, however, is present in 9 out of 10 subjects in the control group, but in only 4 of the disorder group. Whisperiness is present in every voice, in varying degrees. We can conclude that the dysperiodicity of FØ that underlies harshness is a major contributor to the high RATEX and DPQEX values of the disorder group; that the pulse-grouping tendency of creakiness (described by Rabiner et al. (8) as 'diplophonia' where 'alternate pulses are more strongly correlated (both in length and amplitude) than adjacent glottal pulses') is the chief factor in the perturbation scores of the control group; and that whisperiness as such does not contribute powerfully to RATEX and DPQEX except perhaps at an extreme scalar degree. If harshness, creakiness and whisperiness are thought of as components of an overall auditory impression of laryngeal inefficiency, then harshness would be the dominant component, followed closely by creakiness, with whisperiness as a more minor ingredient. The three components would have to be further weighted not only for their respective scalar degrees but also for their degree of intermittency of presence in the voice. The correlation of such a set of weighted perceptual components with, for instance, the GRBAS scale for evaluating hoarseness (G = 'grade', R = 'rough', B = 'breathy', A = 'asthenic', and S = 'strained') proposed in Hirano (9), remains to be investigated, though R seems fairly likely to be an amalgam of harshness and creakiness and B a fairly direct counterpart of whisperiness.

TABLE I

Fundamental frequency measurement by parallel processing (PGR) for speech disorder subjects and a control group, for perturbation (RATEX and DPQEX) and median FØ, with voice type (1-6 = scalar degrees of perceptual rating, H = "harshness", W = "whisperiness", C = "creakiness", V = "voice", i = 'intermittently present')

SPEECH DISORDER GROUP

Disorder type	Voice type	RATEX %	DPQEX %	Median FØ
Dysphonia	5H5WV	74.95	42.77	126.60
Down's Syndrome	5H5W13CV	64.14	40.05	125.00
Dysphonia	4H5WV	57.36	32.18	153.80
Dysphonia	2W2CV	38.94	18.35	92.59
Dysphonia	4H4W12CV	34.51	26.58	126.60
Cerebral Palsy	14H2WV	34.13	22.32	137.00
Cerebral Palsy	4H2W12CV	28.23	22.98	156.20
Down's Syndrome	3H3WV	25.47	19.16	156.20
Deafness	3WV	17.33	12.00	133.30
Cerebral Palsy	3WV	16.36	10.55	151.50
	means	39.14	24.66	135.88

CONTROL GROUP

Speaker	Voice type	RATEX %	DPQEX %	Median FØ
1.	3W3CV	41.60	20.94	92.59
2.	2W3CV	40.63	24.30	117.60
3.	3W12CV	35.06	23.10	116.90
4.	2W12CV	30.95	18.88	116.30
5.	3W12CV	30.91	19.15	113.60
6.	1W2CV	29.98	13.15	116.30
7.	3W3CV	29.12	21.27	103.10
8.	3WV	26.54	19.88	103.10
9.	1W12CV	24.49	16.05	123.50
10.	2W13CV	23.68	14.59	119.00
	means	31.29	19.53	112.20

It is perhaps noteworthy that the two voices with the smallest perturbation values in Table I are to be found in the disorder group. Both these voices showed an audible boost in laryngeal muscle tension, and it may well be that slight increases in myodynamic parameter values improve the efficiency of vibration of the vocal folds.

An important conclusion emerges from the degree of laryngeal inefficiency revealed in the perturbation figures in Table I for the control group. If the voice types recorded in the control

TABLE II

A comparison of fundamental frequency measurements made by parallel processing (PGR), inverse filtering (SIF) and cepstral processing (API) of a dysphonic speaker with extreme FØ perturbation, and a normal control group speaker with moderate FØ perturbation

Speaker	Voice type	Duration of recording	Algorithm	Data-window shift factor	Number of data points	RATEX %	DPQEX %	Median FØ
Dysphonic speaker	5H5WV	60.00 secs	PGR	8 msec	1848	74.95	42.77	126.60
			SIF	8 msec	1607	35.93	12.89	112.00
			API	8 msec	578	28.75	11.61	122.00
Normal speaker	1W12CV	35.33 secs	PGR	8 msec	2581	24.49	16.05	123.50
			SIF	8 msec	2989	14.00	4.79	116.00
			API	8 msec	2373	10.55	4.34	127.00

group are representative of adult males in the general population (and experience in the Edinburgh project strongly indicates that they are), then, as suggested above, inefficiency of laryngeal vibration in voice is the norm not the exception. This means that designers of automatic pitch detection algorithms, in any successful application that involves an identification of personal phonatory quality, will be obliged to address directly the analysis problems caused by quasi-dysphonic perturbations of fundamental frequency in normal voices.

Table II shows a comparison of the intonational and perturbational results obtained from the three pitch detection algorithms, parallel processing (PGR), inverse filtering (SIF) and cepstral processing (API), for a dysphonic speaker with extreme perturbational values and a normal control group speaker with moderate perturbational values, of similar median FØ.

It is evident from Table II that all three algorithms, not surprisingly, perform worse on a severely dysphonic voice than on a normal voice. The particular causes of difficulty in this dysphonic voice lie in the extreme degree of inter-harmonic spectral noise from the whisperiness component, and the drastically wide range of FØ jitter from the harshness component. A major difference in the performance of the three algorithms on the dysphonic voice can be seen in the relative numbers of segments identified as voiced (within the selected range of 50 to 250 Hz), the cepstral method (API) markedly under-reporting voiced segments compared with the other two methods. Over-reporting voiced segments characterises the performance of the inverse filtering method (SIF) on the normal voice. Further analysis is necessary, but the relative tendencies of the three algorithms to make voiced-to-unvoiced and unvoiced-to-voiced errors are broadly in agreement with the findings of Rabiner et al. (8).

The most important difference in the performance of the three algorithms, from the perspective of a pitch detection facility in a clinical application, is to be seen in the very considerable smoothing imposed on the perturbation parameters RATEX and DPQEX by the inverse filtering and cepstral methods, compared with the more perturbation-revealing parallel processing method.

This work was supported by the Medical Research Council (Grant No. G978/1192).

CONCLUSION

Clinical interests in speech signal processing suffer a tension between two opposing requirements in fundamental frequency analysis - the need to register the smoothed trend line of FØ relevant to intonation, and the need to track the momentary, detailed, perturbatory excursion of FØ from that trend line, which contributes to personal phonatory quality. In choosing a general-purpose pitch detection algorithm of reasonable speed, resolution and accuracy, a tradeoff has to be accepted between the two needs. The parallel processing algorithm developed by Gold and Rabiner (2) seems a preferable choice over inverse filtering and cepstral processing methods for such an application.

REFERENCES

(1) J.D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. Audio Electroacoust., vol. AU-20, pp. 367-377, Dec. 1972.

(2) B. Gold and L. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J.Acoust.Soc.Amer., vol. 46, pp. 442-448, Aug. 1969.

(3) A. M. Noll, "Cepstrum Pitch Determination," J.Acoust.Soc.Amer. vol. 41, pp. 293-309, Feb. 1967.

(4) J. Laver and R. Hanson, "Describing the Normal Voice," in J. Darby (ed.) Speech Evaluation in Psychiatry. Grune and Stratton, pp. 51-78:1981.

(5) Y. Koike, H. Takahashi, and T. Calcaterra, "Acoustic Measures for Detecting Laryngeal Pathology," Acta Otolaryng., vol. 84, pp. 105-117, 1977.

(6) P. Lieberman, "Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathological Larynges," J.Acoust.Soc.Amer. vol. 35, pp. 344-353, Mar. 1963.

(7) M.H.L. Hecker and E.J. Kreul, "Descriptions of the Speech of Patients with Cancer of the Vocal Folds," J.Acous.Soc.Amer. vol. 49, pp. 1275-1282, 1971.

(8) L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms," IEEE Trans.Acoust., Speech and Signal Process, vol..ASSP-24, pp. 399-417, Oct. 1976.

(9) M. Hirano, Clinical Examination of Voice. New York: Springer-Verlag, 1981.

APPENDIX 4

AUTOMATIC ANALYSIS OF WAVEFORM PERTURBATIONS IN CONNECTED SPEECH

Steven M. Hiller, John Laver and Janet Mackenzie

ABSTRACT

Details of an algorithm for the automatic acoustic measurement of waveform perturbations in connected speech are presented. A number of measures of perturbations are defined. Results are reported for the application of the algorithm and the perturbation measures to normal voices and a pathological voice, and discussion is offered of the role of the system in screening voices for potential laryngeal pathology.

The automatic analysis of waveform perturbations in connected speech is an extension of a longstanding research interest in the Phonetics Laboratory in the topic of voice quality (Laver 1967, 1968, 1974, 1975, 1979, 1980; Laver & Hanson 1981; Laver, Wirz, Mackenzie & Hiller 1981, 1982; Laver, Hiller & Hanson 1982). Laver (1980) was an early attempt at providing a comprehensive account of perceptual and physiological aspects of normal voice quality, with some preliminary discussion of acoustic aspects. In a recent three-year project ('Vocal Profiles of Speech Disorders' Medical Research Council Grant No. 9781192N, 1979-82), a research team in the Laboratory developed, from this initial base, a perceptual coding system for describing both normal and pathological voice quality. The system was called 'Vocal Profile Analysis', and has now been taught to some 200 speech therapists in a number of different countries. A preliminary account of the system was given in Laver, Wirz, Mackenzie & Hiller (1981), and a full version, supported by illustrative cassette tapes of pathological voices, will be available soon in Laver, Wirz, Mackenzie & Hiller (1984). Now, in a second three-year project ('Acoustic Analysis of Voice Features' MRC Grant No. 8207136N, 1982-85), we are beginning to explore in more detail an acoustic method for characterizing the pathological voice, developing speech signal-processing programs for use on the Laboratory's computer facilities.

This article is a progress report on acoustic and computing aspects of this second MRC project. A companion article (Mackenzie, Laver and Hiller 1983) in this volume reports on anatomical and mechanical aspects of structural pathologies of the vocal folds, and their consequences for perturbatory details of the laryngeal waveform. The project is directed by John Laver; Steve Hiller is responsible for computing aspects, and has written all the computer programs discussed below. Janet Mackenzie is responsible for the speech pathology work. Another member of the project is Robert Hanson, who is a Visiting Senior Scientist from Bell Laboratories, Indian Hills, Chicago: his role is to visit the project each year and advise on signal processing and acoustics.

OBJECTIVES

The broad objective of the project is to explore the feasibility of an automatic acoustic screening system for the early detection of laryngeal pathology. Our first goal is to find acoustic

(Edinburgh University Department of Linguistics, Work in Progress, 16, 40-69, 1983)

parameters, such as dysperiodicity of the fundamental frequency of the laryngeal waveform, which can be used to differentiate the healthy population from those with laryngeal pathologies that perturb the laryngeal waveform. Our later objective is to try to differentiate between the various pathologies of the larynx, initially at a descriptive level, and then possibly from a more diagnostic point of view, on the basis of different degrees and types of waveform perturbations (and other anomalies, such as inter-harmonic spectral noise from incomplete glottal closure due to growths on the vocal folds, paralysis of the vocal folds, etc.). Our third objective is to differentiate between stages of progression, either of a given disease, or of rehabilitative improvement. Even the first of these goals poses considerable difficulties. This is true for various reasons - not the least of which is the fact that almost all current speech signal processing programs available today have inbuilt assumptions that are biased towards the normal model of speech. The more one moves towards abnormal pathology, the more these assumptions are violated, and the less effective the signal processing programs very often become. One of the benefits of working in this area, though, is precisely that these discontinuities (and some continuities) between the normal model and the model we need to develop for the abnormal are highlighted. There is also an important sense in which the study of abnormal malfunction throws light on normal function.

If it is socially important to develop a method of screening the general population for such states as early laryngeal cancer, it is perhaps worthwhile asking the question 'Why choose an automatic acoustic method?' - rather than, say, a perceptual, auditory method, or a physiological method such as electrolaryngography (Fourcin 1974). A number of comments can be offered in reply to this question. Firstly, the provision of an acoustic facility allows an objectivity that a solely auditory approach cannot reliably match. Secondly, as an instrumental technique, an acoustic facility (like physiological facilities) provides a permanent written record which can be repeatedly consulted at leisure, copied for communication purposes, and which allows a detailed quantification of the material analysed. Thirdly, an acoustic facility involves a recording technique that is easily portable, easily used in clinical and other environments, and one which is completely non-invasive. It is a technique that is relatively familiar and unfrightening to patients, and the technology for recording is cheap both in capital and recurrent terms. Because of the portability of acoustic recordings, the analysis facility can be remote from the recording facility in both time and space. This allows a single analysis facility, in some central location, to service a large number of varyingly distant clinics. There are, however, a number of disadvantages to an acoustic facility of this sort. Acoustic signals are inherently contaminable by environmental noise in a way that is less true of physiological signals from such techniques as electrolaryngography. In addition, the remoteness of a central analysis facility brings into consideration factors of communication-links and turn-round time that are less relevant to the technology of local physiological analysis. If an automatic acoustic analysis facility were to be proved feasible for clinical application, then favourable financial criteria come into play. Tape recording facilities are already widespread in hospitals, and the possibility of a single, remote analysis facility minimizes the overall financial outlay, compared with the cost of equipping a wide range of clinics with stand-alone physiological instrumentation. However, a sensible eventual policy

might be to combine the advantages of the two complementary approaches, with a central acoustic facility and local physiological facilities.

An alternative approach would be to adopt local physiological instrumentation and combine it with methods of local acoustic analysis which could be developed for use with microcomputers within each clinic. The one problem with this alternative solution is that, given the currently limited capacity and speed of microcomputers, initial data-acquisition would have to be achieved by special-purpose hardware. Once such combinations of microcomputer plus special-purpose hardware became available, or the speed and capacity of microcomputers increased sufficiently, then the equipment could also be used interactively with the patient as a clinical instrument of assessment and rehabilitation. It is taken for granted that all these approaches combine instrumental techniques with auditory observation by the therapist concerned.

INTONATION VERSUS PERTURBATION

From now on, it will be convenient to concentrate on the role and measurement of just one aspect of speech, that of fundamental frequency (FO).

On close inspection, the succession of pitch periods in voiced speech does not show a perfectly smoothly-changing sequence of durational values, in connected speech. In even the healthiest of voices, the duration of each successive pitch period tends to vary, randomly, from the general trend-line discernible through a sequence of such periods. The trend-line represents the intonational contour, and the local deviations of individual periods from the smooth trend-line, as a perturbation of this trend, are perceived in terms of an auditorily 'rough' phonatory quality. The more dysphonic a voice, the greater is the degree of such perturbation, and the greater is the degree of perceived 'roughness'. One of the problems in choosing a suitable method for the automatic detection of the duration of pitch periods in the acoustic waveform is that there is often then a tension between two quite different needs: the need to establish the smoothed trend which represents the intonational contour, versus the need to register as accurately as possible the momentary deviations (or 'excursions') of individual periods from this smoothed trend, representing phonatory quality. Most pitch period extraction algorithms involve a good deal of smoothing in their inherent design, and as such are well-suited to gathering intonational data. There are very few algorithms available that are capable of tracking the exact durations, cycle by cycle, of the perturbed train of periods that is characteristic of not only dysphonic, pathological voices, but also of many types of normal voices.

The present project is interested in both sorts of data, intonational and perturbational. The algorithm we chose was a parallel-processing method working in the time domain, devised originally by Gold and Rabiner (1969). It was chosen in the light of criteria emerging from comparative studies of a number of pitch period detection algorithms (Rabiner, Cheng, Rosenberg, and McGonegal 1976; Laver, Hiller and Hanson 1982). The Gold and Rabiner method was felt suitable for the project's needs in that it can work on connected speech from both male and female speakers, is resistant

to poor signal-to-noise ratios from recordings in hospital environments, as well as being resistant to interharmonic spectral noise, and retains accuracy of period duration estimation in conditions of fairly acute waveform perturbation in both fundamental frequency ('jitter') and intensity ('shimmer'). Steve Hiller has written a version of the Gold and Rabiner algorithm, and we have developed a number of automatic measures of waveform perturbation. These will be described in turn.

1. AUTOMATIC PITCH PERIOD ESTIMATION SYSTEM

1.0. INTRODUCTION

The basic scheme of the parallel processor, as a very fast program able to be implemented on a general purpose computer, has been described by Rabiner and Schafer (1978:136) as follows:

1. Initial processing of speech signal creates a number of impulse trains which retain the periodicity of the original signal and discard features which are irrelevant to the pitch detection process.

2. This processing permits very simple pitch detectors to be used to estimate the periodicity of each impulse train.

3. The estimates of these simple pitch period detectors are logically combined to infer the period of each laryngeal cycle in the speech waveform.

The idea of parallelism in period detection is that the outputs of a number of simple parallel measures of periodicity for a given speech segment are the inputs to a sophisticated majority logic measure which determines the segment's official pitch period. Gold and Rabiner (1969) suggested that parallelism, as implemented in an automatic pitch period estimator, may simulate the visual observations of a human examining a speech waveform for periodicity.

1.1. THE ALGORITHM

A block diagram of the parallel processor is shown in Figure 1 (adapted from Gold and Rabiner, 1969). The input speech is low-pass filtered to reduce formant information and then processed to produce several functions representing different aspects of periodicity in the waveform. A simple pitch period detector is then applied to each function to determine the periodicity displayed by that function. The various measures of periodicity derived from the functions are then combined in a sophisticated manner to determine the most likely pitch period for the input speech. In addition, processes are required for determining the presence of speech (i.e., discrimination between speech and silence) as well as the likelihood of the resultant pitch period representing a voiced or voiceless segment. The general structure of the program follows the more elaborate version of Gold and Rabiner's (1969) parallel processor in order to accommodate the widest variety of voice types. In the present implementation, the program completes the parallel processing of a given window of speech data and then the window is shifted forward in time to try to capture the next pitch period.

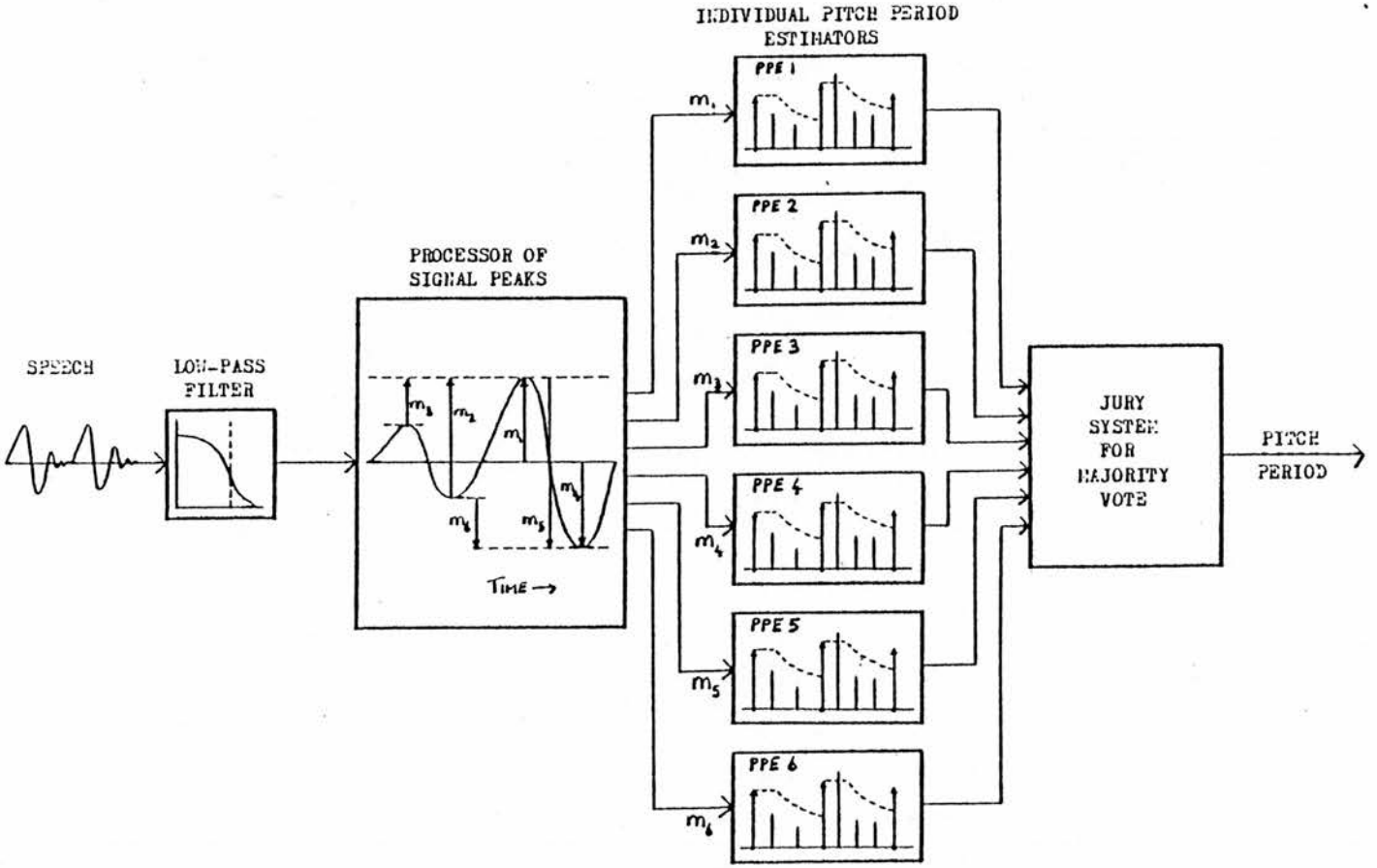


FIGURE 1. Block diagram of the parallel processor for pitch period estimation (adapted from Gold and Rabiner 1969).

1.1.1. Low-Pass Filtering

The input speech signal is low-pass filtered to produce a signal which has been spectrally shaped to contain mostly fundamental frequency information, thus simplifying the period extraction task. In the present system, the low-pass filtering is completed prior to the digitization process by an analog filter. The filter is a Butterworth type which produces a -24 dB/octave rolloff beyond a specified stop band frequency. The cutoff frequency is set to 400 Hz for male voices and 600 Hz for females. This filter also acts as an anti-aliasing filter to prevent spectral distortions during sampling.

1.1.2. Sampling rate

At present, the low-pass filtered signals are digitized at a sampling rate of 10 KHz, as suggested by Gold and Rabiner (1969), thus providing the resolution of pitch periods to within .1 msec. This appears to be a reasonable resolution for typical male fundamental frequencies but increased sampling rates may be required for the higher fundamental frequencies of females and children (Horii, 1979). The digitized signal is then filed for further signal processing.

1.1.3. Silence detection

The pitch period estimation begins by determining the presence of speech within a given window of input data. The silence detection technique is a simple one described by Gold (1964), in which the segment of data is searched for two samples which exceed a pre-determined 'silence' threshold. If the threshold is exceeded then the remainder of the estimation is completed, otherwise the pitch period result is set to zero and the next frame of data is processed. The silence detection threshold is determined interactively for each voice sample by calculating the peak intensity level of the background noise presented in each tape recording. Gold and Rabiner (1969) noted that the parallel processor worked well in low signal-to-noise ratio conditions. This point has been supported for a number of voice samples recorded in rather noisy clinical environments in which good pitch period estimation was possible.

1.1.4. Processing of signal peaks

If speech is present, then the smoothed speech is examined for the presence of "peaks and valleys" (i.e., maxima and minima) which represent periodic behavior in the waveform. Several measures of amplitude are calculated as each valley and peak is located. The amplitude measurement scheme is displayed in Figure 1. This scheme uses six amplitude measurements, which were defined by Rabiner and Schafer (1978, 137) as follows:

1. m1(n): An impulse equal to the peak amplitude occurs at the location of each peak.

2. m2(n): An impulse equal to the difference between the peak amplitude and the preceding valley amplitude occurs at each peak.

3. m3(n): An impulse equal to the difference between the peak amplitude and the preceding peak amplitude occurs at each peak. (If this difference is negative the impulse is set to zero.)

4. m4(n): An impulse equal to the negative of the amplitude at a valley occurs at each valley.

5. m5(n): An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding peak occurs at each valley.

6. m6(n): An impulse equal to the negative of the amplitude at a valley plus the amplitude at the preceding local minimum occurs at each valley. (If this difference is negative the impulse is set equal to zero.)

The use of six different measures of waveform characteristics is designed to cover a range of different types of waveform, varying from a simple sinusoid to a signal composed of a weak fundamental component with a strong second harmonic. Each type of peak and valley measurement produces an impulse train made up of positive impulses representing the amplitudes and locations of the measurements.

1.1.5. Pitch period estimation of the peaks

Each impulse train is evaluated for periodicity by a peak detecting circuit based on an exponential decay function (Gold, 1962). Figure 1 demonstrates the basic operation of this exponential circuit. Following the detection of a possible pitch period marker, the circuit is reset and held for a blanking interval during which no detection occurs. After the blanking interval, the circuit begins to decay. The decay continues until an impulse of sufficient amplitude exceeds the decay threshold, and then is once again reset. In this manner, possible pitch period information is stored and extraneous data discarded. The decay behavior of the exponential circuit (i.e., blanking time and decay rate) is dependent upon local pitch period trends in order that reasonable limits are set for the detection of the next period.

1.1.6. Final computation of the pitch period

For each analysis interval, the peak detecting circuit produces six estimates of the pitch period, one for each of the six impulse trains. These estimates of periodicity are combined with the two most recent sets of estimates from the six parallel pitch period detectors. The final determination of the pitch period is based on a comparison of all the estimates. The estimate with the greatest level of agreement among the six immediate candidates is declared the official pitch period for the speech segment. It should be noted that this method of calculating pitch period causes the loss of some period information at the onset of phonation.

1.1.7. Voiced/voiceless decision

Gold (1964) described the technique used for determining whether the chosen pitch period represents a voiced segment of speech. Voiced/voiceless decisions are determined from the level of agreement between the chosen pitch period estimate and the other period

measures. For voiced speech, the agreement level will be high since each simple detector represents redundant information concerning the periodic behavior of the waveform. There is a lack of redundancy associated with noisy voiceless speech and therefore a low level of agreement for any pitch period estimate. A voiced/voiceless decision threshold can be determined from the distributions of the agreements calculated for voiced and voiceless speech (Gold, 1964).

1.2. ANALYSIS CONDITIONS FOR OBTAINING MICROPERTURBATORY DATA

Since the main objective of the present research is the capture of valid cycle-to-cycle perturbation information, a number of analysis conditions linked to the pitch period estimation process need to be considered. The general approach behind our implementation of the parallel processor is to apply the system to an interval of speech data, accept the last pitch period within an analysis interval detected by the exponential decay system as the representative period, and then shift the window forwards to include the next pitch period. The analysis conditions of most importance to the system are thus the nature of the analysis interval (the analysis 'window'), the shifting of the window, and the waveform feature to be used as a pitch period marker.

1.2.1. Analysis interval conditions

Each pitch period estimation is completed on a segment of filtered speech data selected by a rectangular analysis window. The interval within the window is set to accommodate the largest probable pitch period to be produced by a given speaker. At present, the analysis interval is set to a duration of 25 msec (40 Hz) for male speakers and 20 msec (50 Hz) for female voices. Given the rather long durations of the analysis interval, it is normal for more than one pitch period to be present in the window at any one occasion of period detection. The program has been designed to produce an estimate of period for the last complete cycle in the window.

1.2.2. Shifting of the analysis window

Cycle-to-cycle data is estimated by shifting the rectangular window along the data in such a way as to try to bring just one new pitch period into the window. A shift of 10 msec (100 samples at 10 KHz sampling rate) would thus be ideal for a steady fundamental frequency of 100 Hz. However, this ideal situation is seldom reached, because, in continuous speech, fundamental frequency is naturally moving up and down, both for intonational reasons and for microperturbatory reasons. The algorithm is therefore accurate, in the estimation of any two adjacent periods, only within a certain band of fundamental frequencies. The limits of this band are set by the size of the shift factor, basically. If one considers the situation where a new cycle is being brought into the window by one application of the shift factor, then the longest new period that can be accurately detected is one which is no longer than the shift factor itself. If it is longer, then the previous cycle, already estimated once, remains the last complete cycle in the window, and is re-reported. Under-shifts thus result in over-reporting. Conversely, the shortest new period that can be accurately detected is one which is, at a minimum, greater than half the shift factor itself. If it is half the duration or shorter, then (assuming that the next cycle has the same period or less) the algorithm effectively

3) The prediction of slope was calculated as follows: let Sn equal the variable shift factor to be evaluated as an optimized attempt to bring in the next pitch period Pn economically and accurately, and Mn equal the median value of the five estimated periods prior to that next period. Sn can be estimated on the basis of the difference between the two most recent median values (Mn - Mn-1), this difference being a measure of the slope of the FO trend as estimated at the appropriate delay for the median (i.e., Pn-3). If the difference is equal to zero (i.e., the projected slope is horizontal), then let the next variable shift Sn equal the previous shift factor Sn-1. Otherwise, the next shift is determined from a straight-line approximation from the last median value which includes a factor for the delay, that is, Sn = MN + 3(Mn - Mn-1).

With this variable shift, inaccuracies will arise only under certain conditions of FO movement (leaving aside the consideration of perturbations for the moment). These inaccuracies occur at any intonational corner - i.e., at any point of departure from a straight-line trend. It can be seen that there are limiting values for accurate measurement in these changing contours, beyond which error is inherent.

Figure 2 displays two hypothetical pitch period contours, rising and falling, to which the variable shifting logic has been applied. Each contour (the solid line) is plotted as pitch period duration (ordinate) versus the order of the pitch period estimated sequentially in time (abscissa). The first six points of each contour are the six most recently measured periods. Point Pa on the abscissa is the next period (of as yet unknown duration) to be estimated relative to the shift factor produced by the variable shifting algorithm for medians Mn-1 and Mn. It can be seen for each contour that the zone within which accurate estimation of the incoming period can be achieved (the octave band represented by the dotted line at point Pa) has values determined by the local short-term FO behavior. In the case of the rising pitch period contour (i.e. falling intonation contour), there is tolerance to change-over points (i.e. falling to rising intonation) and no tolerance for rising accelerations of period duration (i.e. increasingly negative intonational slope). For the falling pitch period contour (i.e. rising intonation), there is tolerance for falling accelerations (i.e. increasingly positive intonational slope) and no tolerance for change-over points (i.e. falling to rising periods, rising to falling intonational contour).

Similar constraints operate for perturbed waveforms, and the underlying assumption of orderliness in the data in the form of a straight line tendency becomes progressively invalid with increased severity of cycle-to-cycle perturbatory differences. There are two major problems in severely perturbed waveforms for a variable shifting mechanism of this sort. Firstly, the projection of the predicted slope of FO can swing wildly, giving values for Sn which take extreme forms and which thus minimize the likelihood of effectively capturing the next true period. Secondly, with contributory adjacent median values differing widely, it is logically possible for negative shifts to occur. In these circumstances, using a variable shift can actually be counterproductive, and can

jumps a cycle and reports the next one as the last in the window. Over-shifting therefore results in under-reporting. Thus, an octave band of accurate FO estimation is provided by a given shift factor - this band demonstrating tolerance to increased FOs and intolerance to decreased FOs, relative to the shift factor. This is perhaps less important if one's interest lies in intonation, but it becomes very relevant if the object of attention is perturbatory behavior, where exact cycle-to-cycle measurement is the goal.

It can be seen that the algorithm retains accuracy of perturbatory tracking only to the extent that the combination of intonational and perturbational movement of FO remains within a frequency-zone whose limits are determined by the shift factor. It is clearly helpful if a shift factor can be chosen, in the examination of a given voice, that relates in duration to some statistical property of the period durations to be found in that voice, to optimize accurate pitch period estimation. The simplest pitch-adaptive strategy would be to set the shift factor to one value for males, another for females, and another for children, on the basis of general values found in these populations. The next step in tuning the shift factor to allow accuracy of pitch period extraction would be to adjust it to some statistic of the individual speaker's typical performance, for example, the mean, median, or mode FO of the habitual speech. Finally, one could try to make the shift factor fully pitch-adaptive, using strategies to change the value of the shift factor dynamically, on the basis of predictions about future short-term period behavior reached from examinations of local past short-term history of FO. These three types of pitch-adaptive strategies will be referred to as sex-specific tuning, speaker-specific fixed tuning, and speaker-specific variable tuning.

All three types of approach were used experimentally in comparing the benefits of fixed and variable settings of the shift factor. For each speaker, we made a preliminary pass through the data, using a sex-specific shift setting of 10 msec (this setting was for male speakers). From this, the median FO was calculated and used to give a fixed shift which was speaker-specific. Alternatively, the sex-specific setting was used as a starting point for processing the speaker's data by means of a variable shift factor. This variable shift was calculated as follows:

1) An assumption was made that there is an underlying orderliness in the train of pitch periods in speech. In the extreme case this would be represented by an FO contour which would be a straight line - level, rising, or falling. Within voices that can be considered to be normal and healthy, microperturbatory excursions can be anticipated to be infrequent, to be small in extent, and to have a normal distribution for size of excursion.

2) What was needed was some means of predicting the slope of the FO trend, from knowledge of recent FO trend behavior. One possibility was to use a moving-average approach to establish the history of recent FO trend. But means are very vulnerable to the influence of single eccentric values. So it was decided to base the prediction of slope of the FO on recent medians. We chose a moving 5-point median.

itself contribute artifactually to high perturbation values. A partial solution, up to moderate perturbation levels, is to set range-limits. We set a range-limit of 40 to 240 Hz for male speakers. When the extrapolated shift fell outside this limit, the calculation was cancelled, and the first-pass speaker-specific value was substituted. At the same time, a flag was set for each occurrence of this out-of-range incident, to keep a measure of how often the range-limits were invoked, and the first-pass speaker-specific shift value substituted.

1.2.3. Pitch period markers

This condition is concerned with the choice of waveform features which yield pitch period markers. Normally, preference is given to the pitch period marker which relates to the positive peak impulse function (see m1 in section 1.1.4 above). The positive peak detector was chosen for two reasons. First, the final computation of the period by majority logic is biased towards the positive peak when comparisons between the various period measures produce equivalent levels of agreement (e.g. in smooth unperturbed segments of voiced speech). Second, the positive peak parameter is the one most directly related to the impulse behavior of the vibrating vocal folds. Further bias towards the positive peak has been added to the system to accommodate small variations in period measurements. It was observed early on that the period durations varied slightly between some of the pitch period detectors for a given pitch period. The slight variations appear to be the result of actual differences for the various features of the low-passed waveform, and perhaps of the effects of the digitizing process. In these cases, it was observed that some other pitch period marker (e.g., m4 - the negative peak measure) had the highest level of agreement even though the positive peak marker was clearly visible and similar in duration. This is the logical consequence of a program which uses past information and redundant features to arrive at a final decision. It was decided, after inspection of typical waveforms, to force the final measure of the period to be the positive peak marker if the difference in duration between some other chosen feature and the positive peak measure was minimal. For the time being, the system is set to choose the positive peak marker if there is a difference of less than or equal to 3 msec between the two period measures. This minimal difference appears to work well for a majority of cases, as will be discussed below. Differences greater than 3 msec are accepted as an indication of perturbed behavior in the waveform and the alternative peak marker duration is stored.

1.3. PERFORMANCE OF THE AUTOMATIC PARALLEL PROCESSOR COMPARED TO VISUAL OBSERVATIONS OF NORMAL SPEECH

It was important to evaluate the performance of the pitch period extraction system when applied to data of known characteristics. In particular, we were concerned with the behavior of the system under the two methods of window shifting (fixed and variable) which we felt would have the greatest effect on accurate perturbation measurement. The following discussion is based on a small pilot study to determine the types of error produced by the automatic system in comparison to visual examinations of speech stimuli.

1.3.1. The pilot study

The automatic pitch period extractor and visual examinations were applied to the stimulus utterance 'A rainbow is a division

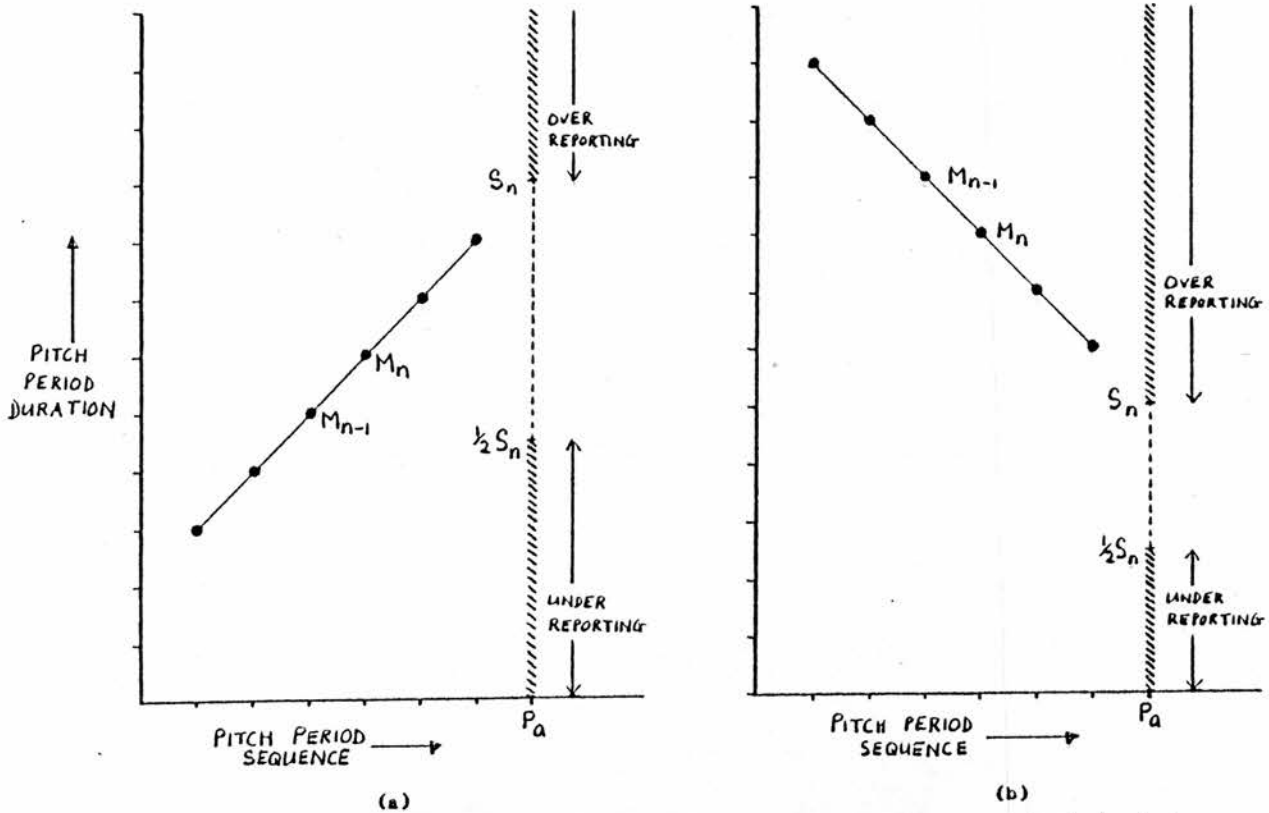


FIGURE 2. Application of the variable shifting algorithm to two hypothetical pitch period sequences (a - rising, b - falling). The two medians Mn and Mn-1, derived from the six most recently measured periods, are used to predict the next period at point Pa. Note how the octave band (at Pa) for accurate period estimation is located relative to recent F0 behavior.

of white light into many beautiful colors'. Tape recordings of the utterance were produced by three normal-speaking male adults (RK, JL, SH). The parallel processor was applied to the data in two manners: 1) shifting of the analysis window by a fixed speaker-specific shift factor based on the median period duration derived on a first-pass analysis of the stimulus and 2) variable shifting using a shift factor based on the median shifting logic presented in section 1.2.2. The output of the automated system was compared with visual examinations of the low-pass filtered versions of the stimuli using a cursor program on the minicomputer's visual display unit. The results of the comparisons are summarized in Tables I and II for the fixed and variable shift conditions for each speaker.

1.3.1.1. Under-reporting/over-shifting; over-reporting/ under-shifting

There is a marginal advantage in these normal voices for the variable shift. In other words, the distribution of FO values for each speaker falls typically within the accuracy span of the shift-setting of the fixed shift, and making the shift-setting pitch-adaptive brings only a small improvement. It is noteworthy that there is an overall low incidence of pitch period over-reporting for each utterance, given the intolerance of the octave band to FOs deviating towards lower frequencies relative to the local FO trend. This result suggests that the intonational behavior evidenced in the utterances was mostly free of decelerating changes from the local FO trends, and that falling intonational contours typically followed more straight-line tendencies. Further research into a more refined mechanism for variable shifting is currently being undertaken, however.

1.3.1.2. Over-reporting due to shimmer factors in sudden low-amplitude values for waveform peaks

Recalling that an exponential decay function is an integral part of the period detection algorithm, when shimmer factors drop the amplitude of waveform peaks below the exponential threshold, the next true peak is usually beyond the shifted window, and the previously reported cycle is treated as the last complete cycle in the window and re-reported. Values for this type of error were low in both the fixed and the variable shift operations, and the differences were negligible. However, this ability of shimmer factors to contribute to jitter data should persuade us, as Askenfelt and Hammarberg (1980, 1981) suggest, to talk of waveform perturbation, rather than of jitter alone.

1.3.1.3. Non-positive pitch period marker

Despite the bias towards the positive peak parameter, occasionally some other aspect of the waveform receives the majority vote. The figures are very low in both cases due to the additional forcing logic for small variations between simple pitch period detector durations.

1.3.1.4. Voiced-to-unvoiced errors

Occasional low levels of agreement between simple period estimates due to perturbations in the waveform result in an improper unvoiced decision relative to the visual estimation of

Subject	RK N=199, CTX=96	JL N=216, CTX=93	SH N=211, CTX=90
Under-reporting/ Over-Shifting	5.0% (10)	4.2% (9)	7.1% (15)
Over-reporting/ Under-Shifting	2.5% (5)	5.1% (11)	1.4% (3)
Over-reporting/ Low Amplitude	1.5% (3)	1.4% (3)	3.3% (7)
Non-positive Peak Detector	3.5% (7)	3.2% (7)	2.4% (5)
Voiced-to- Unvoiced Error	0.5% (1)	0.5% (1)	1.4% (3)

TABLE I

Errors in automatic pitch period estimation, using a FIXED shift factor, relative to visual estimation, in three normal male voices.

Subject	RK N=198, CTX=96	JL N=216, CTX=93	SH N=225, CTX=90
Under-reporting/ Over-Shifting	4.5% (9)	3.7% (8)	3.6% (8)
Over-reporting/ Under-Shifting	2.5% (5)	3.2% (7)	2.6% (6)
Over-reporting/ Low Amplitude	1.5% (3)	2.3% (5)	3.1% (7)
Non-positive Peak Detector	3.0% (6)	3.2% (7)	3.1% (7)
Voiced-to- Unvoiced Error	0.0% (0)	0.5% (1)	1.8% (4)

TABLE II

Errors in automatic pitch period estimation, using a VARIABLE shift factor, relative to visual estimation, in three normal male voices.

the waveform. The number of voiced-to-unvoiced errors is very low for the data, and supports the findings of Rabiner et al. (1976).

2. PERTURBATION ALGORITHM

2.0. INTRODUCTION

The design of the algorithm for calculating microperturbatory behavior was primarily based on the nature of fundamental frequency contours extracted from continuous speech. The fundamental frequency curves of continuous speech represent modulations of FO associated with intonational aspects of the utterance as well as short-term microperturbations of FO correlated with efficient use of laryngeal vibration. Continuous speech also introduces the influence of segmental performance into the FO contour, such as pauses, and voicing onsets/offsets, and the effects of stop closures, nasality, etc. In addition, the process of pitch period estimation sometimes produces artifacts in the contour through incorrect period estimations. The primary choice of perturbation algorithm was based on the need for a system which provided intonational information in the form of the underlying smooth curve of the raw pitch periods; this curve being a useful baseline from which to measure the variation of the raw pitch periods from local smoothed behavior. Secondarily, the system had to be able to cope with the segmental and artifactual features evidenced in FO contours of continuous speech.

2.1. THE ALGORITHM

The raw FO curve extracted by the parallel processor is passed through a non-linear smoother to produce a contour equivalent to the smoothed underlying trend of the data. The non-linear smoother was implemented as a running digital filter which enables the determination of excursion behavior of the raw FOs from the local smoothed output of the filter. The smoothed FOs and their associated excursions are statistically evaluated for intonational and microperturbatory measures.

2.1.1. The trend line

The trend line underlying the raw FO curve is constructed by a non-linear smoother presented by Rabiner, Sambur, and Schmidt (1975). A non-linear smoother has advantages over more conventional linear smoothers (e.g., running average) which tend to smear sharp discontinuities present in speech signals as well as being affected by gross errors in the contour. A non-linear smoother was chosen since we wanted to preserve realistic discontinuities present in FO contours - these discontinuities representing transitions from voiced to voiceless states and vice versa - while smoothing microperturbatory roughness and gross pitch period estimation errors. The non-linear smoother implemented is a combination of running median filter plus a Hanning window.

The median filter serves to preserve sharp discontinuities in the FO contour, where the desirable discontinuities must be of a minimum critical duration. Rabiner et al. (1975) noted two significant characteristics of the median filter. First, the size of the median filter is based on the minimum duration which defines an acceptable discontinuity. In the present system, we are con-

cerned with discontinuities which represent transitions from voiced to voiceless states, and vice versa, evidenced in FO contours of continuous speech. Voiced (i.e., greater than 0 Hz) and voiceless (i.e., equal to 0 Hz) segments of a FO contour were operationally defined as those segments consisting of three or more sequential FOs of either state. Therefore, a median filter with a duration of five samples is required to preserve discontinuities of three samples or more. Second, the median filter inherently smooths out sharp discontinuities in the signal which are shorter than the minimum acceptable duration. In our system, very short discontinuities are considered to be gross errors in pitch period extraction and, as a result of the operational definition for segments, one and two point discontinuities are smoothed out by the 5-point median filter. The advantage of this second characteristic of the median filter is that large errors do not affect the surrounding calculations of the trend line.

A Hanning window is used as a linear smoother to filter out the less sharp noise components evidenced in speech signals. In the present research, the noise components represent microperturbatory movements in the raw FO contour. A 3-point Hanning window is used in the non-linear smoother as recommended by Rabiner et al. (1975).

The combination of a 5-point median filter and a 3-point Hanning window results in a filtering delay of 3 points for the non-linear smoother. Rabiner et al. (1975) noted the need for additional logic for determining the beginning and ending points of the output data which are lost due to the filter delays. The primary concern of the present research is quantifying the perturbation behavior in the FO contour and therefore the onset and offset data are not included as part of the perturbation data.

2.2.2. Excursions

Excursions represent the deviations of the raw FOs from the equivalent smoothed values produced by the non-linear smoother. The use of excursion measures from a smoothed trend has been presented in Koike (1973), Kitajima, Tanabe and Isshiki (1975), Davis (1976), Kitajima and Gould (1976), Koike, Takahashi and Calcaterra (1977), and Laver, Hiller and Hanson (1982). Excursions are measured relative to a smoothed trend line in order that slow-moving modulations (e.g., vibrato) and intonational movements of FO are excluded from contributing to perturbation parameters.

An excursion is derived for each output of the non-linear smoother and defined as the difference between the raw FO and its equivalent smoothed FO. Each excursion is stored in four formats: 1) signed excursion in Hz - the difference between raw and smoothed FO in units of Hz with the algebraic sign retained, 2) signed excursion in percent - the ratio of the signed excursion in Hz to its associated smoothed FO multiplied by 100, 3) magnitude excursion in Hz - the absolute value of the signed excursion in Hz, and 4) magnitude excursion in percent - the absolute value of the signed excursion in percent. The signed and magnitude excursions in percent are used to normalize excursion measures extracted from varying FO levels evidenced within a sample of continuous speech of a single speaker as well as between different speakers. Excursion measures are not calculated for voiceless segments of FO contours and regions of very short discontinuities. In the latter case, a flag is set denoting each instance of a short discontinuity.

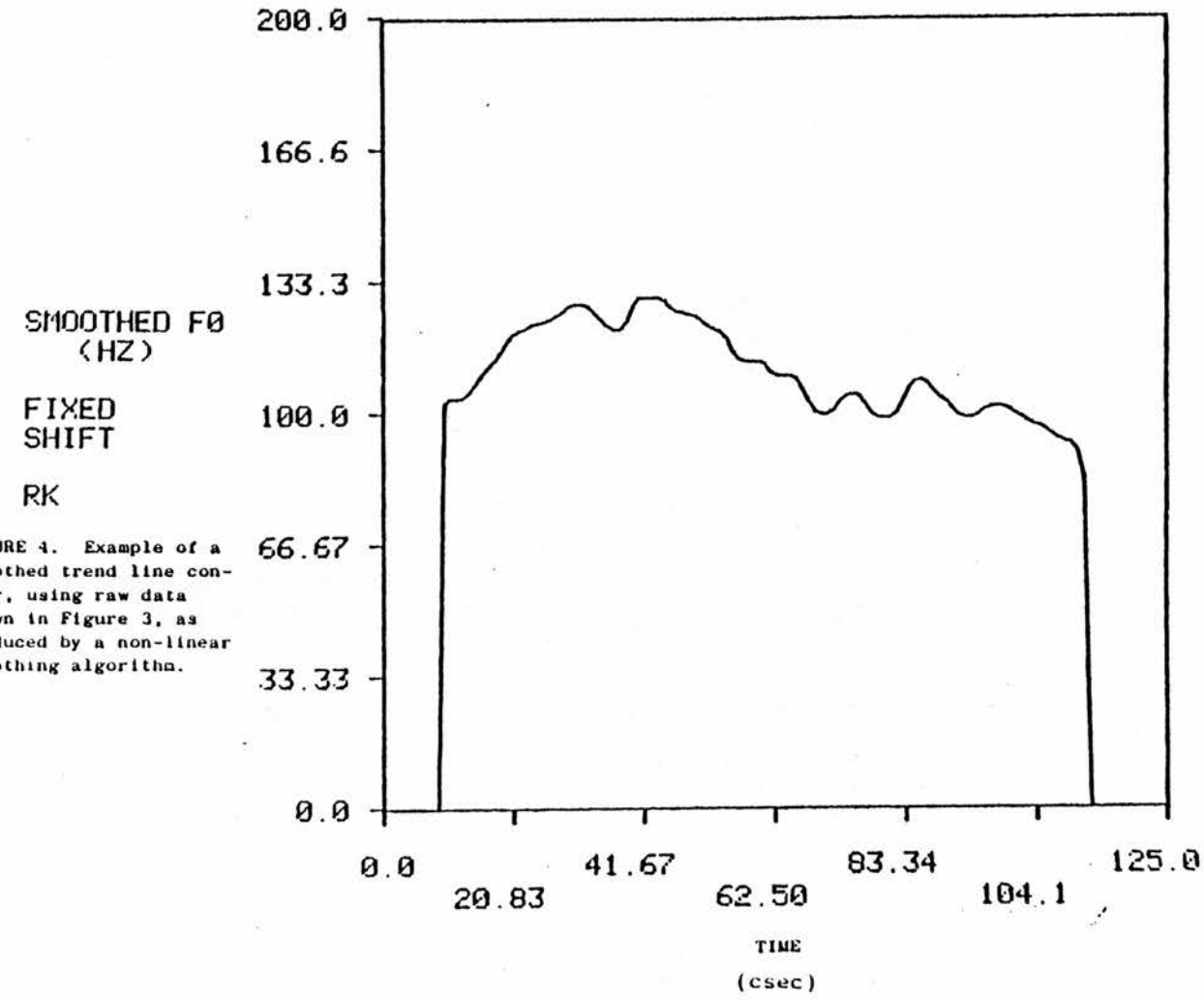


FIGURE 4. Example of a smoothed trend line contour, using raw data shown in Figure 3, as produced by a non-linear smoothing algorithm.

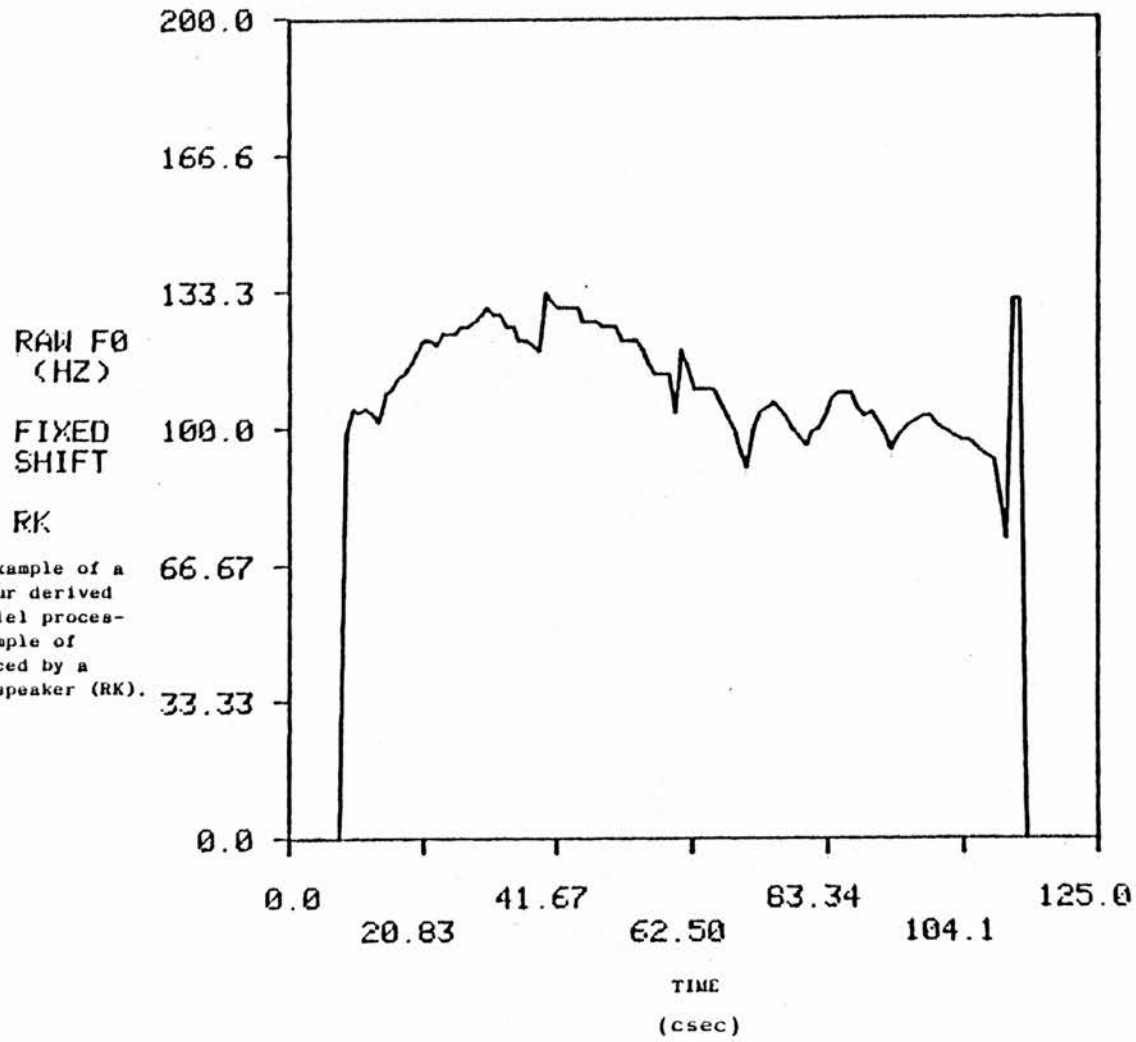


FIGURE 3. Example of a raw F0 contour derived by the parallel processor for a sample of speech produced by a normal male speaker (RK).

Figures 3 and 4 demonstrate an example of the smoothing of a raw FO contour by the non-linear smoother. Figure 3 is the raw FO contour derived by the parallel pitch period estimator for a small section of the stimulus utterance produced by a normal speaker (RK). This contour is characterised by a normal range of FO values for a male speaker as well as small irregularities in the contour typical of normal phonation. Figure 4 is the equivalent FO contour produced by the smoothing process. This smoothed trend line retains the over-all intonational features of the original raw FO contour but the small irregularities have been removed, thus making the trend line a useful base from which to measure the irregularities.

2.2.3. Perturbation measures

Several statistical measures are determined for each FO contour which describe the magnitude, distribution, and frequency of micro-perturbatory behavior in each sample.

AVEX - The average magnitude of the excursions present in each FO contour, described in units of Hz or percent.

SDEVEX - The standard deviation of the distribution of the excursions present in each FO contour, described in units of Hz or percent.

RATEX - The Rate of Excursions is the percentage of points in the sample where a magnitude excursion in percent is equal to or greater than a pre-set threshold. RATEX is adapted from the "Pitch Perturbation Quotient" (PPQ) of Koike, Takahashi, and Calcaterra (1977) - RATEX differs from PPQ in that a non-linear smoother is used to produce a smoothed FO trend rather than the moving average approach used to calculate PPQ. The non-linear smoother preserves major features of the FO contour while smoothing out noisy and anomalous components in the contour. RATEX is based on magnitude excursions in percent in order to normalize excursion measures calculated for varying FOs evidenced within and between speakers' phonations. The pre-set threshold is used to quantify the number of significant perturbations in any given speech sample (similar to Lieberman's (1963) minimum threshold for his "Perturbation Factor"). The pre-set threshold is set to 3%, because even in the healthiest voice, uttering a monotone vowel, the successive pitch periods typically show approximately 2% frequency jitter, in a normal distribution (Hanson, 1978). A 3% threshold allows us to discount this factor. Thus RATEX reflects the incidence of significant excursions in the sample.

DPF - The Directional Perturbation Factor which has been adapted from Hecker and Kreul (1971). DPF is the percentage of changes of algebraic sign calculated for differences between adjacent raw FO measures (thus not based on a smoothed trend line). A 3% threshold for the magnitude of the difference between adjacent FOs is also included in this measure to exclude the normal distribution of FO differences.

ANOMALIES - This category includes both short discontinuities, as defined earlier, and anomalous FOs outside the pre-set range for acceptable frequencies (i.e., 40-240 Hz for males, 75-450 Hz for

females). All such anomalies are rejected from the perturbation calculations, but their occurrence flagged.

OUT OF RANGE - The total number of occasions when, in the variable shift calculation, the projected value of the incoming period fell outside the pre-set limits for acceptable FO values.

2.2.4. Intonation measures

Several measures of overall intonational behavior are calculated for each utterance based on the smoothed trend line. These measures include the mean, median, mode, and standard deviation of the FO distribution, limited to the pre-set limits for acceptable frequencies.

2.2.5. Application of the perturbation algorithm

The perturbation algorithm was applied to the FO contours extracted from three normal voices (RK, JL, SH) and one very pathological voice (MA2/RIE12), each speaker having produced the test sentence 'A rainbow is a division of white light into many beautiful colours'. Tables III and IV present the resultant perturbation and intonational measures for each of the voices - Table III contains data derived via a speaker-specific fixed shift factor and Table IV displays data derived via a speaker-specific variable shift factor. Inspecting the perturbation measures AVEX, SDEVEX, RATEX, and DPQEX, it can be seen that a clear separation of the pathological speaker from the normal speakers exists. Similar results were noted for the ANOMALIES and OUT OF RANGE measures between the pathological and normal speakers.

Figures 5, 6, 7 and 8 display examples of FO intonational and perturbational distributions produced by the perturbation programs for one normal speaker (RK) and one pathological speaker (MA2/RIE12). All data presented in these figures were derived via the speaker-specific fixed shift method of pitch period estimation. Figures 5 and 6, for the normal and pathological speakers respectively, are histograms of the long term FO intonational behavior based on the smoothed trend line output of the non-linear smoother. The FO histogram of the normal speaker (mean = 109.2 Hz) in Figure 5 shows a distribution which is more normally distributed, narrower and more peaked compared to the FO distribution of the pathological speaker (mean = 126.2 Hz). The FO histogram of the pathological speaker shows a bimodal distribution. Figures 7 and 8 are histograms of the short term perturbational data based on the signed magnitude of the excursions in Hz for the two speakers. These two figures also demonstrate substantial differences for the phonatory behavior of the two speakers with a much narrower and more peaked distribution for the normal speaker compared to the pathological speaker. The differences between the two perturbational distributions are reflected in the greater AVEX, SDEVEX, and RATEX measures of the pathological speaker compared to the normal speaker.

CONCLUSION

Having developed a successful pitch detection algorithm, and plausible measures of perturbation, the next stage of the project is to apply these to an extensive set of voices. These fall into two categories. The first involves the recorded voices of patients

Subject	RK N=228, CTX=46	JL N=300, CTX=43	SH N=283, CTX=40	MA2/RIE12 N=848, CTX=70
AVEX	3.57 Hz 3.24 %	3.38 Hz 3.65 %	4.85 Hz 3.48 %	16.20 Hz 12.26 %
SDEVEX	11.46 Hz 10.48 %	7.47 Hz 4.12 %	15.08 Hz 4.86 %	24.42 Hz 18.02 %
RATEX	14.04 %	21.33 %	20.44 %	52.12 %
DPQEX	8.41 %	20.13 %	15.04 %	34.01 %
ANOMALIES	5	2	6	42
F0 MEAN	107.20 Hz	105.60 Hz	115.20 Hz	126.20 Hz
F0 MEDIAN	99.00 Hz	104.40 Hz	113.00 Hz	128.30 Hz
F0 SD	13.11 Hz	13.64 Hz	18.71 Hz	37.57 Hz

TABLE III

Automatic F0 perturbation analysis for three normal male voices (RK, JL, SH) and one dysphonic male voice (MA2/RIE 12), using a FIXED shift factor.

Subject	RK N=225	JL N=243	SH N=306	MA2/RIE12 N=726
AVEX	2.66 Hz 2.24 %	3.44 Hz 3.43 %	4.54 Hz 3.46 %	14.61 Hz 10.62 %
SDEVEX	7.82 Hz 6.55 %	10.18 Hz 10.20 %	12.24 Hz 10.74 %	22.81 Hz 16.15 %
RATEX	15.56 %	20.48 %	23.20 %	48.35 %
DPQEX	8.70 %	17.85 %	13.11 %	35.47 %
ANOMALIES	1	1	10	35
OUT OF RANGE	2	0	5	43
F0 MEAN	109.20 Hz	107.50 Hz	114.50 Hz	126.40 Hz
F0 MEDIAN	99.20 Hz	105.40 Hz	117.30 Hz	128.30 Hz
F0 SD	15.47 Hz	13.45 Hz	19.60 Hz	37.35 Hz

TABLE IV

Automatic F0 perturbation analysis for three normal male voices (RK, JL, SH) and one dysphonic male voice (MA2/RIE 12), using a VARIABLE shift factor.

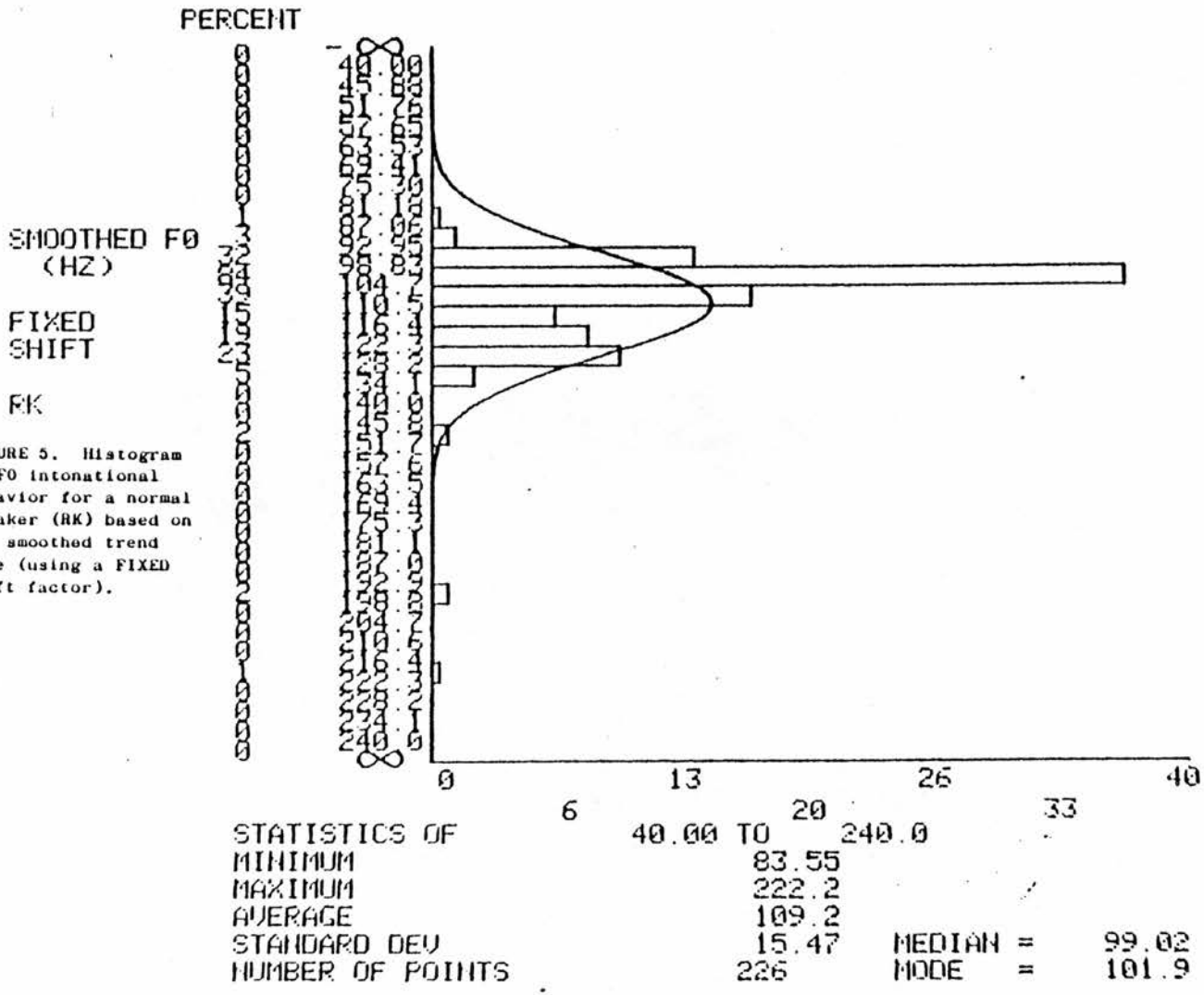


FIGURE 5. Histogram of F0 intonational behavior for a normal speaker (RK) based on the smoothed trend line (using a FIXED shift factor).

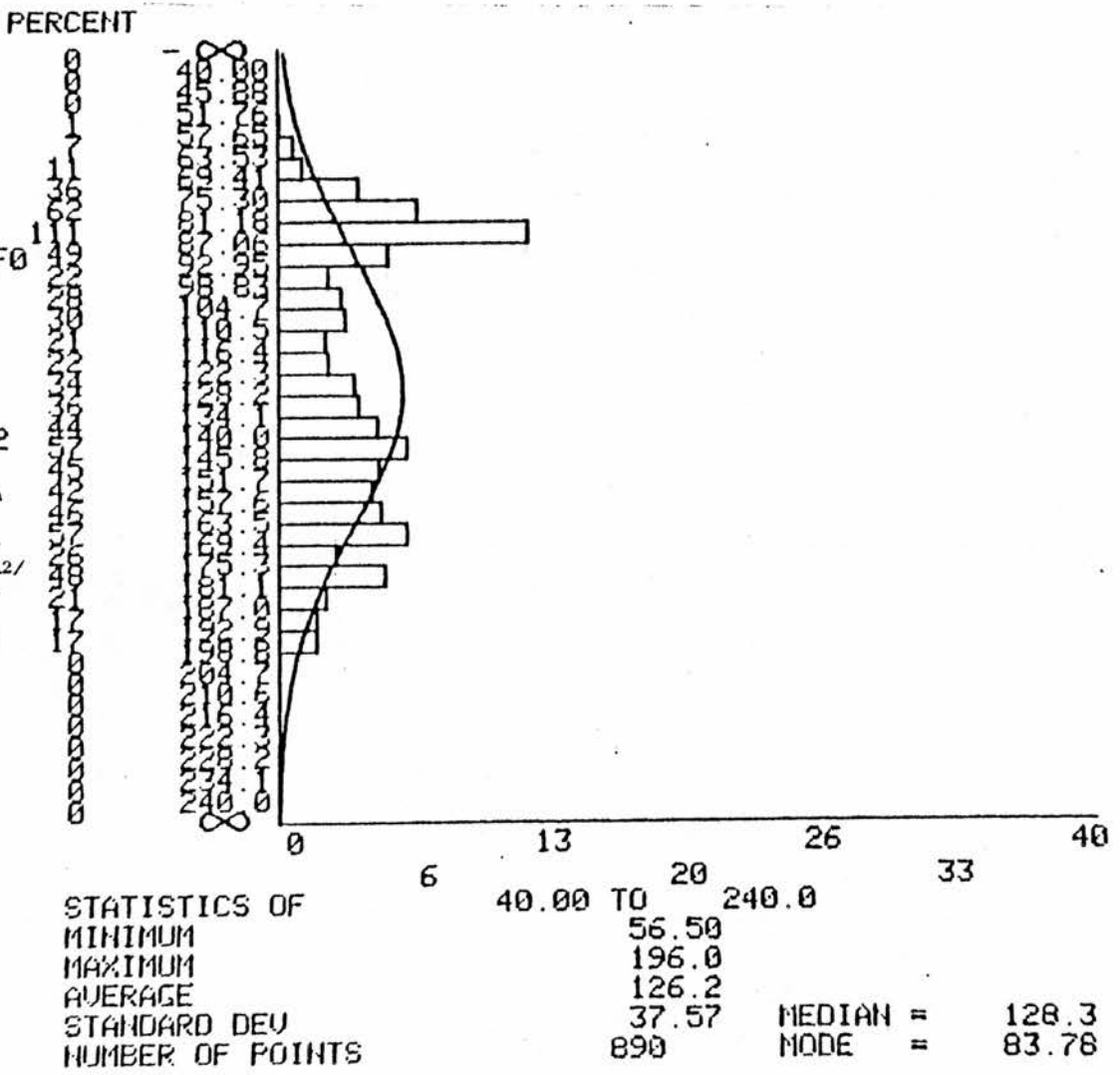


FIGURE 6. Histogram of F0 intonational behavior for a pathological speaker (MA2/RIE12), based on the smoothed trend line (using a FIXED shift factor).

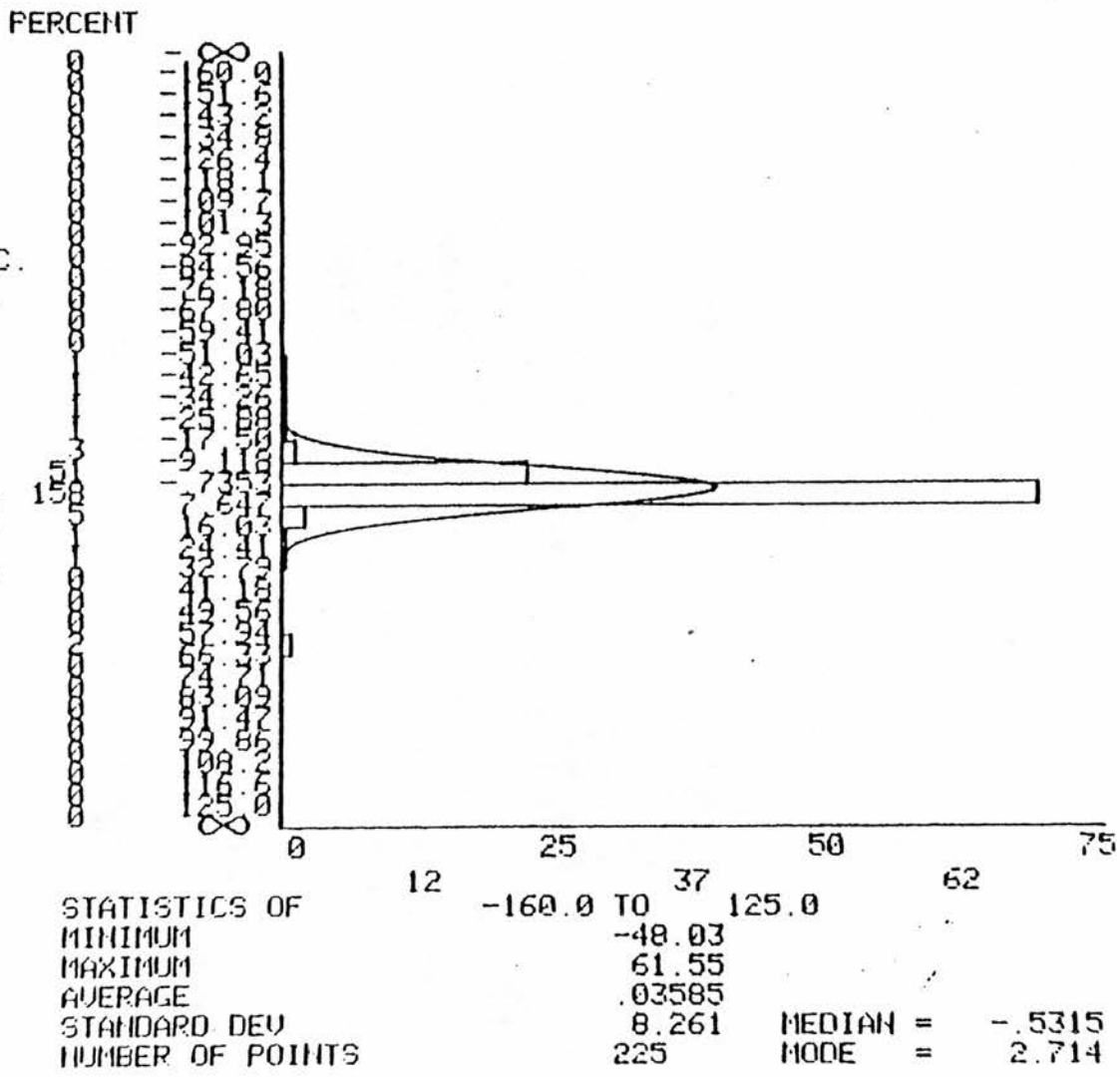


FIGURE 7. Histogram of signed excursions in Hz for a normal speaker (RK) using a FIXED shift factor.

from voice clinics, for which a wide range of diagnostic information about their vocal pathology will be made available. The two main collaborating institutions for this part of the project are the Otolaryngology Departments of the Radcliffe Infirmary, Oxford, and the Royal Infirmary, Edinburgh. The project will seek to correlate acoustic perturbational data with the type and degree of pathology present, as discussed in more detail in Mackenzie, Laver and Miller (1983: see this volume). The second category consists of a control group of some hundred voices of each sex. A fairly clear picture is available of most of the acoustic characteristics of the normal voice (Laver and Hanson, 1981), but this does not yet include a full knowledge of typical ranges of perturbation in the healthy voice. This is needed to establish the phonatory norm from which pathological voices can be held to deviate.

The hypothesis underlying the work of the project is that increasing perturbation, beyond a threshold yet to be established, reflects increasingly severe pathology. This hypothesis will obviously have to be refined, and the range of perturbation which characterises stages of different pathologies will have to be made more specific, but as a preliminary conceptual step it seems profitable to distinguish between two general levels of perturbation. The first of these is the range of perturbation that characterises the normal, healthy larynx: we can refer to perturbation in this range as being "microperturbation". The second is the range of perturbation that characterises the unquestionably pathological larynx: we can call this more extreme type "macroperturbation". As an initial estimate, the threshold for passing from microperturbation to macroperturbation possibly lies somewhere in the range between 30 to 40% RATEX, with an associated AVEX of 10% or more and SDEVEX of 15% or more - i.e., where roughly between a third or more of all individual periods in phonation deviate substantially and variably from the local smoothed trend line.

Given that our interest is in screening the general population for potential laryngeal pathology, rather than only in quantifying the phonatory consequence of unquestionable pathology, it is the border zone towards the end of the microperturbatory range, up to the threshold of definitely pathological macroperturbation, that attracts our attention. This is the zone of perturbation where, within the frame of reference of a screening system, an individual subject can be held to be 'at risk', as indicated in Figure 9. This 'risk zone' is where early signs of pathology will surface, we speculate. It may well be that the phonation of a given speaker found to be in the risk zone will be one where the relatively high degree of microperturbation shown is due to the dysperiodic symptoms of a particular habitual but healthy phonation type, such as creaky voice (vocal fry), rather than of pathology. But false alarms of that sort are the price one pays for the benefit of a screening system designed to catch symptoms of vocal pathology as early as possible. A major part of our empirical research will consist of tuning the boundaries of the risk zone as far as possible to reduce false alarms and maximize the early detection of laryngeal pathology. This tuning process will include the investigation of the differential power of the perturbation measures to distinguish between the populations of normal and pathological speakers.

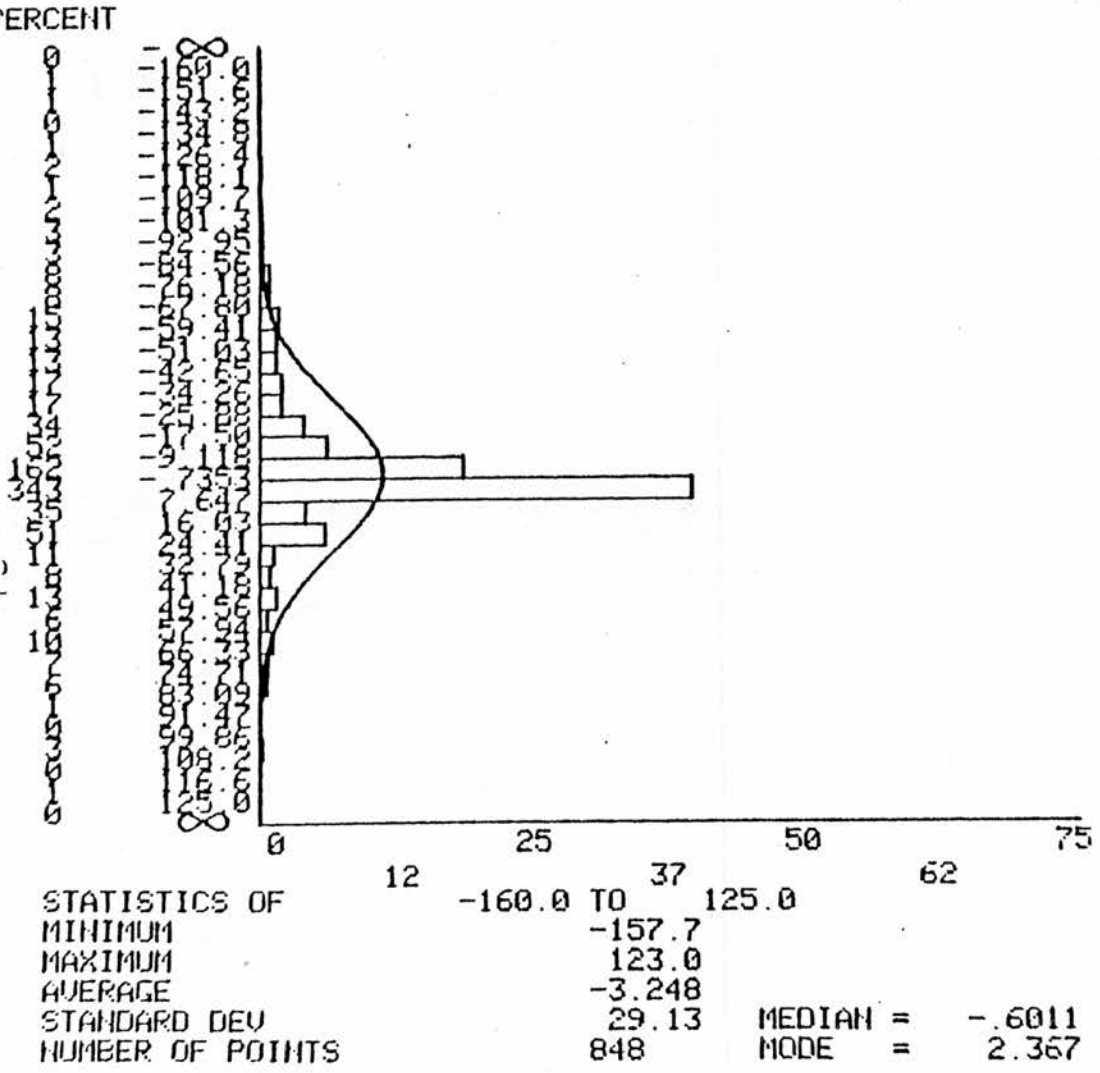


FIGURE 8. Histogram of signed excursions in Hz for a pathological speaker (MA2/RIE12) using a FIXED shift factor.

STATISTICS OF	-160.0 TO 125.0		
MINIMUM	-157.7		
MAXIMUM	123.0		
AVERAGE	-3.248		
STANDARD DEV	29.13	MEDIAN =	-.6011
NUMBER OF POINTS	848	MODE =	2.367

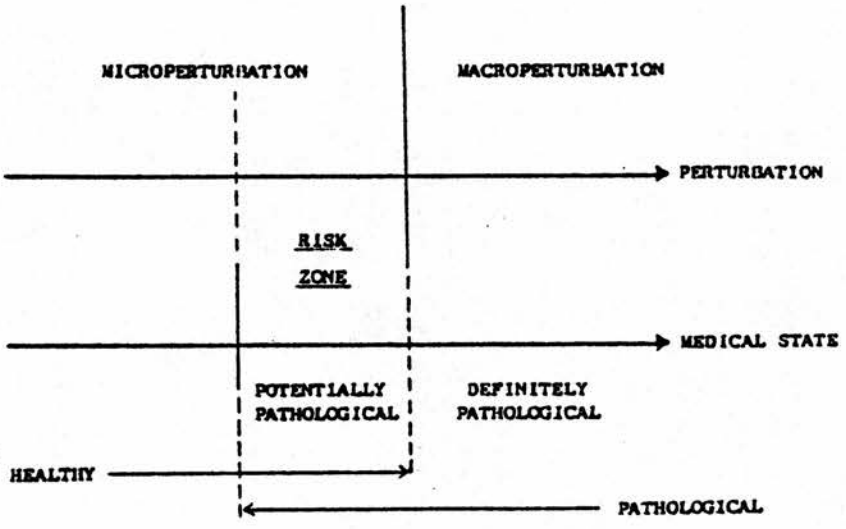


FIGURE 9. A schematic diagram of the relationship between waveform perturbation and vocal fold pathology.

REFERENCES

Askenfelt, A. and Hammarberg, B. (1980) 'Speech waveform perturbation analysis'. Speech Transmission Laboratory Quarterly Progress and Status Report, 4, 40-49.

---------- (1981) 'Speech waveform perturbation analysis revisited'. Speech Transmission Laboratory Quarterly Progress and Status Report, 4, 49-68.

Davis, S. B. (1976) 'Computer evaluation of laryngeal pathology based on inverse filtering of speech'. Speech Communication Research Laboratory Monograph, 13.

Fourcin, A. J. (1974) 'Laryngographic examination of vocal fold vibration'. In B. Wyke (ed.), Ventilatory and Phonatory Control Mechanisms. London: Oxford University Press, 3-5-333.

Gold, B. (1962) 'Computer program for pitch extraction'. J. Acoust. Soc. Am., 34, 442-448.

---------- (1964) 'Note on buzz-hiss detection'. J. Acoust. Soc. Am., 36, 1659-1661.

Gold, B. and Rabiner, L. R. (1969) 'Parallel processing techniques for estimating pitch periods of speech in the time domain'. J. Acoust. Soc. Am., 46, 442-448.

Hanson, R. J. (1978) 'A two-state model of FO control'. J. Acoust. Soc. Am., 64, 543-544.

Hecker, M. and Kreul, E. (1971) 'Descriptions of the speech of patients with cancer of the vocal folds. Part 1: Measures of fundamental frequency'. J. Acoust. Soc. Am., 49, 1275-1282.

Horii, Y. (1979) 'Fundamental frequency perturbation observed in sustained phonation'. J. Speech and Hearing Res., 22, 5-19.

Kitajima, K. and Gould, W. J. (1976) 'Vocal shimmer in sustained phonations of normal and pathological voices'. Annals. Otol., 85, 377-381.

Kitajima, K., Tanabe, M., and Isshiki, N. (1975) 'Pitch perturbations in normal and pathological voice'. Studia Phon., 9, 25-32.

Koike, Y. (1973) 'Application of some acoustic measures for the evaluation of laryngeal dysfunction'. Studia Phon., 7, 17-23.

Koike, Y., Takahashi, H., and Calcaterra, T. C. (1977) 'Acoustic measures for detecting laryngeal pathology'. Acta Otolar., 84, 105-117.

Laver, J. (1967) 'The synthesis of components in voice quality'. Proceedings of the VI International Congress of Phonetic Sciences, 523-535, Czechoslovak Academy of Sciences, Prague.

Laver, J. (1968) 'Voice quality and indexical information'. Brit. J. Disorders Comm., 3, 43-54.

---------- (1974) 'Labels for voices'. J. Inter'l. Phonetic Assoc., 4, 62-75.

---------- (1975) Individual features in voice quality. Doctoral dissertation, University of Edinburgh.

---------- (1979) Voice Quality : a Classified Bibliography. Amsterdam: John Benjamins B.V.

---------- (1980) The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press.

Laver, J. and Hanson, R. J. (1981) 'Describing the normal voice'. In J. Darby (ed.), Speech Evaluation in Psychiatry. New York: Grune & Stratton, 51-78.

Laver, J., Hiller, S. M., and Hanson, R. J. (1982) 'Comparative performance of pitch detection algorithms on dysphonic voices'. Proceedings of IEEE Conference on Acoust., Speech, and Signal Proc., 192-195.

Laver, J., Wirz, S., Mackenzie, J. and Hiller, S. M. (1981) 'The perceptual protocol for the analysis of vocal profiles'. Work in Progress, Department of Linguistics, Edinburgh University, No. 14:139-155.

Laver, J., Wirz, S., Mackenzie, J. and Hiller, S. M. (1982) Vocal profiles of speech disorders. Final Report on MRC Grant No. 9781192N, University of Edinburgh.

Laver, J., Wirz, S., Mackenzie, J. and Hiller, S. M. (forthcoming 1984) Vocal Profiles. Cambridge University Press.

Lieberman, P. (1961) 'Perturbations in vocal pitch'. J. Acoust. Soc. Am., 33, 597-603.

---------- (1963) 'Some acoustic measures of the fundamental frequency periodicity of normal and pathological larynges'. J. Acoust. Soc. Am., 23, 361-363.

Mackenzie, J., Laver, J., and Hiller, S. M. (1983) 'Structural pathologies of the vocal folds and phonation'. Work in Progress, Department of Linguistics, Edinburgh University, No. 16:80-116.

Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., and McGonegal, C. A. (1976) 'A comparative performance study of several pitch detection algorithms'. IEEE Trans. Acoust., Speech and Signal Proc., ASSP-22, 552-557.

Rabiner, L. R., Sambur, M. R., and Schmidt, C.E. (1975) 'Applications of a non-linear smoothing algorithm to speech processing'. IEEE Trans. Acoust., Speech and Signal Proc., ASSP-22, 552-557.

Rabiner, L. R. and Schafer, R. W., (1978) Digital Processing of Speech Signals. New Jersey: Prentice-Hall, Inc.

APPENDIX 5

DURATIONAL ASPECTS OF LONG-TERM MEASUREMENTS OF FUNDAMENTAL FREQUENCY PERTURBATIONS IN CONNECTED SPEECH

Steven Hiller, John Laver and Janet Mackenzie

An important consideration in many applications of automatic speaker characterization is establishing the minimum sample duration for long-term acoustic parameters to reach stability. Stability is here understood to refer to the extraction of parameters in such a fashion that they genuinely characterize the speaker rather than the message content of the speech sample. The characterization of a speaker by long-term acoustic parameters is the first step in a variety of speech pattern recognition experiments. For example, certain types of speaker recognition studies are designed to be text-independent with the major restriction on the stimulus materials being their overall length (Rosenberg 1976). Similarly, the description of characteristic voice quality in normal and pathological speakers requires a speech sample long enough to permit the abstraction of long-term average features of overall quality from the fluctuating values of short-term segmental performance (Laver, Wirz, Mackenzie and Hiller 1981).

One easily available acoustic parameter for characterizing speakers is the long-term behavior of fundamental frequency (Atal 1976). Most studies of fundamental frequency are based on long-term feature averaging. Fundamental frequency (FO) is one of a variety of speech parameters for which Markel, Oshika and Gray (1977) suggest that the concept of a long-term average value is relevant, despite the lack of a "true" mean or variance in FO data in real speech due to non-random factors such as the declination effect in intonation. One major aspect of long-term features of FO is the overall intonational behavior of a given speech sample - included in this category are such detailed measures as mean, median, mode, standard deviation, skewness, kurtosis, etc. For long-term intonational FO statistics, Nolan (1983) noted a general finding within the relevant literature that within-speaker variation was minimized for durations of approximately one minute. Steffan-Batog, Jassem and Gruzska-Koscielak (1970) found that 50 seconds of read text was enough to produce a regular distribution for long-term averaging of FO. Green (1972) suggested that segments of speech as short as 15 seconds would flatten the short-term variations of pitch in samples of conversational speech. Distributional convergence of FO statistics occurred for a sample duration of approximately 60 seconds for read scripts analyzed by Horii (1975). Mead (1974) reported that an optimal duration of 75 seconds was useful for a speaker recognition task using unconstrained speech, but durations as short as 30 seconds approximated longer duration results in cases where only short samples of speech were available for analysis. Markel and Davis (1979) noted that durations between 40 and 70 seconds were sufficient to show convergence to a stable long-term average FO for linguistically-unconstrained speech.

The comments above concern intonational behavior, where the frequency value of the general trend-line through the FO data is of greater relevance than very local irregularities in adjacent periods. But of course, on close inspection, the succession of pitch periods making up the intonational contour of voiced speech does not show a perfectly smoothly-changing sequence of duration values. The duration of each successive pitch period tends to vary randomly

(Edinburgh University Department of Linguistics, Work in Progress, 17, 59-76, 1983)

periods from the local smoothed behavior of the contour. It was felt that a non-linear smoother was appropriate for processing segmental and artifactual features evidenced in machine-analyzed FO contours of continuous speech (Hiller et al. 1983).

As output, the system provides distributional histograms and statistical data such as the mean, median and standard deviation of the smoothed FO data, which are used to represent long-term intonational behavior. Similar statistical data is provided for the measured differences between the actual and smoothed curves; these differences are referred to as excursions of the actual FO periods from their smoothed equivalents in the trend line. The excursions of FO are the basic units for determining long-term perturbational features for each speech sample. Long-term features of perturbation include the mean and standard deviation of the excursions as well as the rate and direction of perturbatory movement as calculated for the entire FO contour. The rate of excursions (RATEX) is the percentage of points in the sample where the magnitude of an excursion in percent is equal to or greater than a pre-set threshold. The pre-set threshold is used to quantify the number of salient perturbations in any given speech sample. A pre-set threshold of 3% was used since a normal speaker, uttering a monotone vowel, evidences an approximate 2% frequency jitter, in a normal distribution (Hanson 1978). The Directional Perturbation Factor (DPF) of the excursions, adapted from Hecker and Kreul (1971), is the percentage of changes in algebraic sign calculated for differences between adjacent actual FO measures. A 3% threshold for the magnitude of the difference between adjacent FOs is also included in this measure to exclude the normal distribution of FO differences.

Subjects, Speech Material and Analysis Procedures

High-quality tape recordings were made during oral reading of the first two paragraphs of "The Rainbow Passage" (Fairbanks 1960) by 20 adults (10 males and 10 females) who had no known history of speech or voice disorders. Nineteen of the speakers were British (Scottish or English) with the remaining speaker being an American. Prior to the recording, each speaker familiarized himself with the passage and was asked to read at a comfortable loudness level. The recorded speech samples were digitized using a PDP 11/40 computer and stored on computer magnetic tape for further processing by a VAX 11/750 computer. For subsequent analyses, the total utterance of each speaker was divided into 5 second segments. Intonational and perturbatory statistics were completed for durations successively incremented by 5 seconds (i.e. 5, 10, 15, ..., 60 seconds) up to 60 seconds or the end of the utterance, whichever came first. Each incremental analysis therefore incorporated information from the previous shorter duration analyses.

Results and Discussion - Male Speakers

Figures 1 and 2 display examples of the results of the present study for two of the 5 acoustic parameters measured as a function of increasing sample duration (incremented every five seconds) for the 10 male speakers. The two parameters are the mean FO in Hz (Figure 1) and the average magnitude of the excursions in percent (Figure 2). The perturbation parameters are measured in percent to normalize for differing levels of FO. The results for each of the 10 male speakers are labelled in each figure as A - J, with each figure representing the parameter on the ordinate versus

from the general intonational trend line discernible through a sequence of pitch periods. These local deviations of individual periods from the smooth intonation are considered perturbations of the FO contour and auditorily perceived as a 'rough' phonatory quality (Hiller, Laver and Mackenzie 1983). Such perturbatory deviations are usually larger in degree and more frequent in speakers suffering from laryngeal pathology than in normal speakers. Because of their role as signals of potential laryngeal disorder, therefore, short-term perturbatory movements of FO have often been measured on a long-term statistical basis for the assessment of such voice disorders (see, for example, Lieberman 1963; Hecker and Kreul 1971; Davis 1979; Askenfelt and Hammarberg 1980, 1981; Laver, Hiller and Hanson 1982). The long-term measures of perturbatory FO behavior typically include the mean, range and rate of occurrence of these perturbations in voiced speech. In most studies of perturbation, FO acoustic parameters are derived from sustained vowel phonations or short sentences. A few studies have attempted to extract data from longer durations of continuous speech, which can be considered a more natural use of phonation (Askenfelt and Hammarberg 1980 1981; Laver et al. 1982).

The literature reported above suggests that intonational values stabilize to a steady mean in speech data whose minimum duration lies somewhere between 40 and 75 seconds. But a corresponding minimum duration for stabilization of perturbational values is not yet well established, and the purpose of this study is to determine the minimum duration of continuous read speech which will stabilize long-term acoustic parameters of perturbation in normal speakers. The opportunity will also be taken to further the study of the minimum duration for stabilization of intonational characteristics. In addition, the relationship between sampling frequency and resolution of fundamental frequency will be discussed in relation to data duration.

Fundamental Frequency Measurement

All FO data were obtained by a computer program which used a modified parallel processing method of peak peaking in the time domain (Gold and Rabiner 1969; Hiller et al. 1983), operating on speech waveforms which were sampled at either 10 or 20 KHz. A pre-processing stage prior to digitization consisted of low-pass filtering the speech data to eliminate most formant information from the waveform (filter cutoff frequencies were set to 600 Hz for males and 800 Hz for females). The digitized speech was processed in parallel through six simple pitch detectors, each examining a different aspect of peak periodicity in the waveform. A sophisticated jury system was applied to the six period estimates to determine the estimate with the highest confidence. A pre-set level of confidence had to be exceeded by the jury vote before the estimate was considered voiced.

A post-processing stage consisted of smoothing the output of the parallel processor by a non-linear smoothing algorithm which combines a 5-point running median with a 3-point Hanning window (Rabiner, Sambur and Schmidt 1975). The choice of this type of smoother was based on the need for a procedure which provided intonational information in the form of the smooth curve derived from the actual periods - this curve being a useful baseline from which to measure the perturbatory variation of the actual pitch

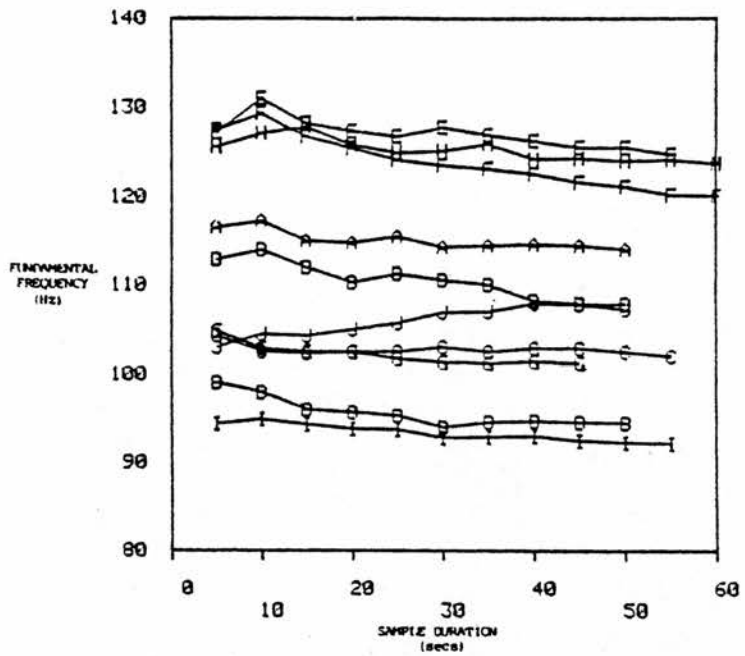


Figure 1 Changes in long-term value of fundamental frequency with increasing sample duration (in cumulative 5 second increments) for 10 normal male speakers (labeled A-J).

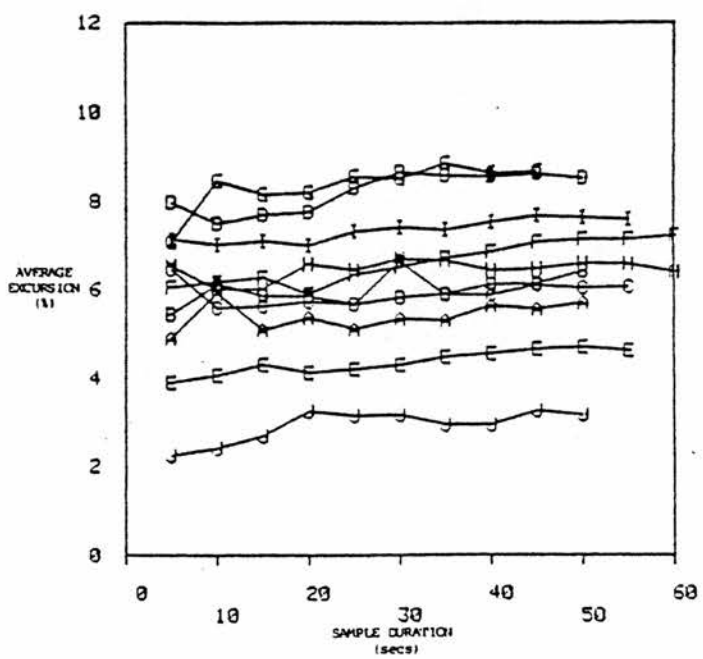


Figure 2 Changes in long-term value of average excursion with increasing sample duration (in cumulative 5 second increments) for 10 normal male speakers (labeled A-J).

increasing duration in seconds on the abscissa. It should be recalled that each 5 second increment on the abscissa includes data from the previous segments up to and including the specified duration. In addition, each duration analyzed for each speaker represents the actual length of the recorded speech, which includes both voiced and unvoiced speech. It was found that the ratio of voiced to unvoiced data points increased linearly as duration was increased for all the male speakers' data. This finding suggests that the male speakers spoke the text with a regular tempo and no intermittent long periods of voicelessness. The average ratio of voiced to unvoiced data points was approximately 66% for male speakers and 70% for females.

The general description for the 10 speakers' durational curves for all 5 parameters is one of instability at the shorter durations below approximately 25 seconds followed by a flattening of each parameter by the end of the spoken text. Visual estimates of the data suggest that 40 seconds of spoken text is a sufficient duration for all 5 parameters to reach stability. The 5 parameters included the standard deviation of F0, the standard deviation of the excursions, and DPF, as well as mean F0 and mean excursion displayed in Figures 1 and 2, respectively. A value of 40 seconds of spoken text agrees broadly with the findings of other researchers, particularly those studies in which a standard text was used as the stimulus material. No large differences were noted between intonational and perturbational measures for overall duration required for parametric stability (comparing, for example, Figures 1 and 2). The general growth patterns for the intonational and perturbational parameters were dissimilar - the intonational measures demonstrating values which decreased with increasing duration while the perturbation measures display increasing values with time. In Figure 1, the mean F0 is seen to lower slightly in frequency as more data is accumulated for most of the speakers. The decrease in mean F0 may be the result of 2 effects: (1) paragraph effects associated with the linguistic structure of read English and (2) progressively decreasing tension (i.e. decreased stiffness) of the vocal folds during continuing oral reading. The notable exception to the downward trend for mean F0 is male speaker J, who demonstrates increased F0 with increased duration of the data. The F0 values for the 10 speakers are spread across a frequency range reported by other researchers for normal male voices. The perturbation measure in Figure 2 (the average magnitude of the excursions) demonstrates the tendency towards increased perturbation values with time which may be correlated with decreased efficiency of phonation with progressive laryngeal fatigue experienced in the speaking of long texts. These growth curves for perturbation measures may provide some useful information about voice pathology, following the precedent of the notion of articulation growth curves for speech discrimination testing in audiology.

Effects of Sampling Resolution on the Time-Domain Analysis of Fundamental Frequency Periods

Following the analysis of the male voices for long-term features of intonation and perturbation, a question arises as to the intrinsic accuracy of the F0 results based on speech data sampled at 10 KHz. A pitch period extractor working in the time domain is initially limited in its measurement accuracy by the effects of temporal quantization due to the sampling resolution. For example, a sampling rate of 10 KHz is equivalent to a resolution

of .0001 sec, thus producing steps of approximately 1 Hz at a 100 Hz FO (i.e. 99.0, 100.0, 101.0 Hz). It can be seen that a difference in measuring resolution occurs between typical male and female fundamental frequencies, and this difference affects the long-term measurement accuracy of FO for these voices (see, for example, Horii 1979). Figure 3 displays the effects of two sampling rates, 10 and 20 KHz, on the resolution of a number of FO levels. The resolution measure is labelled Just Noticeable Difference of FO (JND FO) and measured as a percentage factor to normalize for the differences in FO levels. The JND FO measure is calculated as the ratio of the absolute difference between the FO level and the next possible FO (this difference is based on the temporal resolution of the given sampling frequency) to the set FO level:

$$\text{JND FO in \%} = \frac{FO_n - FO_{n-1}}{FO_n} * 100$$

where FO_n is the frequency corresponding to a given period. For example, at a FO level of 100 Hz, and at a 10 KHz sampling rate, the next possible step up in frequency is approximately 101.0 Hz which is equivalent to a JND FO of 1%. The JND FO represents the minimum frequency movement (i.e. perturbation) which can be measured for local small variations in FO for a given sampling resolution. There are two curves for each of the sampling rates in Figure 3 - a step up in frequency is represented by the upper curve and the lower curve represents a step down in FO. It is evident that for both sampling rates increasing FO level directly controls increasing JND FO. That is, for any given sampling rate, sampling resolution becomes poorer as fundamental frequency is increased. The 'factor of 2' relationship between the two sampling rates is obviously preserved for the resultant JND FOs at each of the FO levels.

The JND FO can be used to determine the minimum acceptable error in FO measurement due to sampling resolution. Hess (1983) reviews psychoacoustic evidence which highlights the acceptable levels of FO measurement associated with sampling resolution. If the extracted frequency data is to be used as part of a speech synthesis system, then the JND for the audition of FO changes is the most sensitive indicator of measurement resolution. For example, Holmes (1973) noted that the quantization of pitch due to sampling resolution was audible even for a JND FO of less than 1%. Hess reported JNDs of FO which were low for synthetic vowels (0.3 to 0.5% for the male FO range) compared to the JNDs derived from sinusoidal tones of equivalent frequency. Higher JNDs (4.0 to 5.0%) were noted for real speech stimuli, the greater scores being attributed to the presence of perturbations in the real speech waveforms as compared to the synthetic speech. For experimental tasks other than synthesis, FO measurements with less accuracy than the JNDs for audition may be acceptable. Hess suggests that the next acceptable level of accuracy could be based on speech production rather than perception, assuming the accuracy of voluntary adjustment of FO in production is poorer than the perceptual JNDs. A review of a number of perturbation studies leads Hess to conclude that "with respect to accuracy, the ear thus outperforms the speech production system by far". The magnitude and occurrence of perturbations in normal speech are great enough to exceed the JNDs for

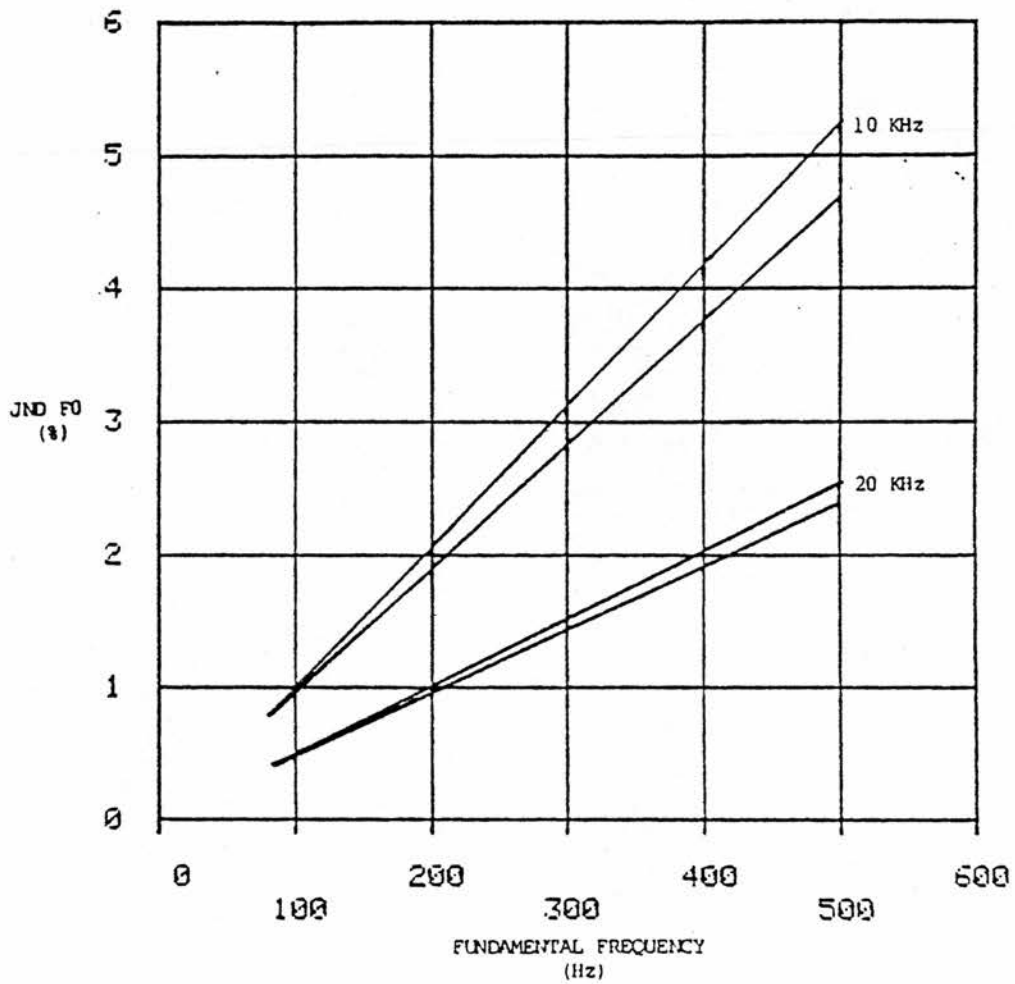


Figure 3 Just Noticeable Difference (JND) of F0 in percent plotted as a function of absolute F0 (Hz).

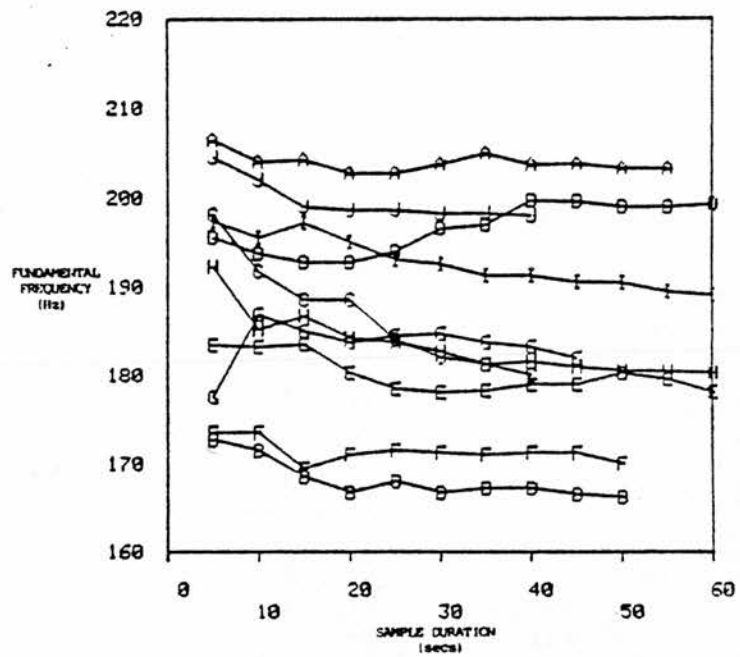


Figure 4 Changes in long-term value of fundamental frequency with increasing sample duration (in cumulative 5 second increments) for 10 normal female speakers (labeled A-J).

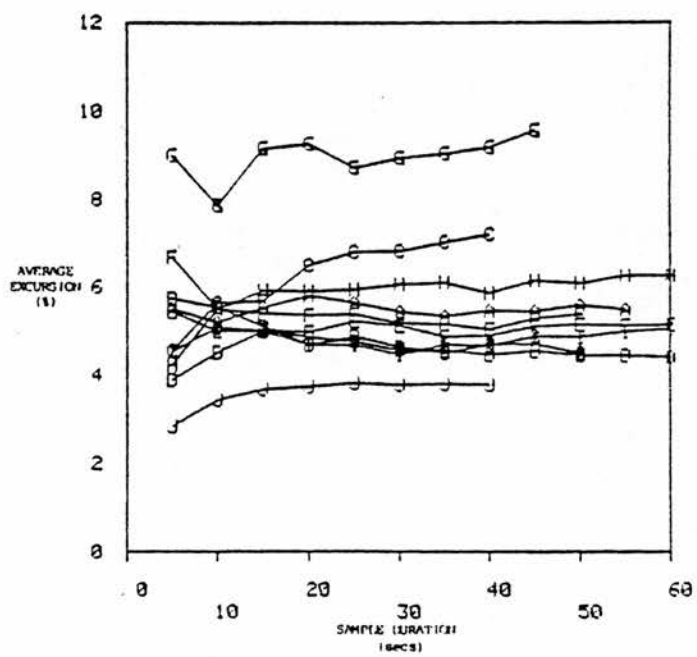


Figure 5 Changes in long-term value of average excursion with increasing sample duration (in cumulative 5 second increments) for 10 normal female speakers (labeled A-J).

FO; the perception of the perturbations being one of phonatory 'roughness' rather than any change of pitch. Finally, the next level of acceptable accuracy of FO resolution could be based on the linguistic relevance of FO changes, for instance, in the case of FO patterns correlated with the perception of stress in English. Hess reports experimental data which suggests that the JND for linguistically-relevant FO changes (presumably in languages other than tone languages) can be as high as 18 to 25%.

In the present study, interest is focused on the accurate measurement of FO in order to characterize speakers. In the first instance, it is reasonable to select a sampling resolution which will quantize FO to a JND level typical of voluntary adjustment of FO production - this level being greater than the JND for the perception of FO changes in speech. However, a greater level of resolution accuracy is required if the perturbations in natural speech are to be measured and related to the perception of phonatory roughness. Thus, a JND FO of less than or equal to 3% would be a reasonable compromise between measurement accuracy associated with speech production and perception. It can be seen in Figure 3 that at the higher frequency levels, a sampling frequency of 20 KHz produces FO measurement accuracy below the JND FO of 3%. Therefore, a sampling rate of 20 KHz would be suitable for quantizing the perturbatory activity of the higher pitches of female speakers and many children.

The use of higher sampling rates does have a number of consequences. First, a much greater amount of digital storage and processing time will be required to analyze long durations of speech data. Second, data recorded from speakers with very high FOs may require very high sampling rates in order to detect the presence of perturbations correlated with the early stages of laryngeal pathology. In addition, there is still the problem that any fixed sampling frequency will yield a differential resolution at different fundamental frequency values, both within the performance of a single speaker and between different speakers. Interpolation of the sampled data (e.g. by parabolic interpolation or upsampling techniques) to increase the apparent sampling rate may be a useful technique for overcoming the overhead associated with increased resolution as well as improving FO measurement accuracy.

Results and Discussion - Female Speakers

In the light of the above comments about sampling frequency and resolution, data for female speakers was sampled at 20 KHz. Figures 4 and 5 present the results for 10 female speakers for the same 2 parameters (mean FO and average magnitude of the excursions) measured as functions of incremented sample duration. Note that the vertical scale of Figure 4 differs from Figure 1 in order to accommodate the female speakers' FO values. In each figure, the parameter value is displayed versus sample duration, and the 10 female speakers labelled A - J. The growth patterns of the parameters derived from the female data are similar in nature to the male results. Early instability of the parameter values is followed by movement towards stability. The long-term duration of 40 seconds for relative parametric stability also applies to the female results. Intonational growth curves demonstrated decreasing values with increasing sample duration though female speaker D produced increased mean FO values with increased duration. The overall values of the FO means for the female speakers are nearly

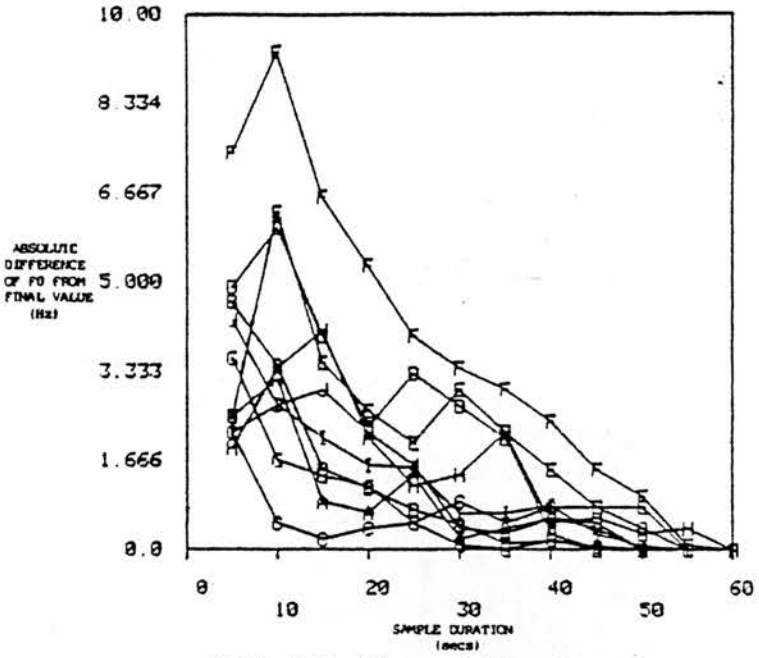


Figure 6 Absolute differences of fundamental frequency from the final long-term value plotted against sample duration (in cumulative 5 second increments) for 10 normal male speakers (labeled A-J).

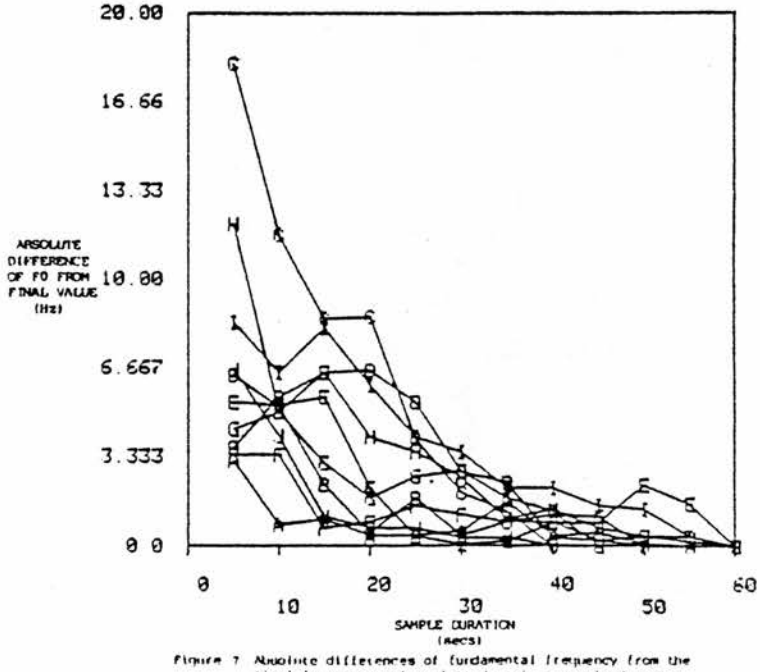


Figure 7 Absolute differences of fundamental frequency from the final long-term value plotted against sample duration (in cumulative 5 second increments) for 10 normal female speakers (labeled A-J).

an octave greater than the male FOs. The growth curves for the perturbation parameters revealed increased values as duration was increased for the female speakers, but not with the same consistency as the male speaker's curves. The resultant perturbation values for the female speakers are similar to the male results which suggests that the perturbation analysis method successfully normalizes for differing levels of FO. One notable result was female speaker G who demonstrated perturbation values which are much higher than the other speakers. This speaker's results may be related to her history of heavy smoking. For all the parameters, the distributions of parametric values were narrower for the female speakers as compared to the male speakers.

Further Assessment of the Durational Data

To assess the durational data, absolute difference curves were derived from the data contained in the original durational curves. A difference curve is composed of values which are the absolute differences between a speaker's final long-term parametric value for a read passage and each cumulative value at each 5 second time increment. Figures 6 and 7 display absolute difference curves for the mean FO parameter for the 10 male speakers and the 10 female speakers. The vertical axis of each figure represents the difference in Hz for each cumulative value at a given time increment (the abscissa) from the final long-term mean FO. It should be noted that the vertical scales in Figures 6 and 7 differ in magnitude due to the larger difference measures of some of the female speakers. The difference curves demonstrate decreasing differences in Hz as the time increments approach the final durational values for FO (which naturally drop to zero for the final durations). As would be expected, the difference curves behave in a similar manner to the original curves in that the difference curves appear to stabilize around the 40 sec increment. One advantage of using the absolute difference curves is that all the parameters demonstrate decreasing values with increasing time thus permitting comparisons between the various parameters. For example, Figures 8 and 9 display the difference curves for the average magnitude of excursions in percent for the 2 groups of speakers (the vertical scales of these 2 figures also differ due to the larger difference values for some of the female speakers).

A further advantage found for the difference curves is that threshold tests can be applied to determine acceptable durational stability for each parameter. A threshold can be defined as some agreed proportion of the final long-term value of a given parameter. It is best expressed in this case as a percentage of the final value, rather than as a single absolute value of FO in Hz, in order to normalize between different speakers. Another consideration is the proportion of members of the group of subjects whose difference curves successfully pass a given threshold. It was decided that 95% of the subject group would form a suitably representative proportion. Thus, the conjunction of the two criteria allows the minimum duration for group-stability to be established.

As a further condition, it was decided that once each difference curve passed a given percentage threshold, it should remain within that band for at least 10 secs further. This condition ensures some stability in a parameter's behavior for a given threshold. The problem of the differing durations of the various speakers' speech samples is partly addressed by this stability condition.

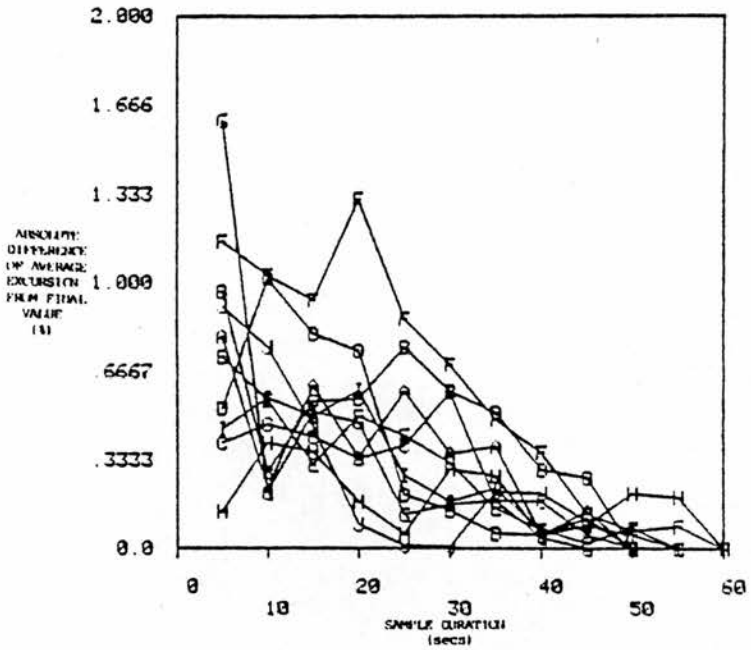


Figure 8 Absolute differences of average excursion from the final long-term value plotted against sample duration (in cumulative 5 second increments) for 10 normal male speakers (labeled A-J).

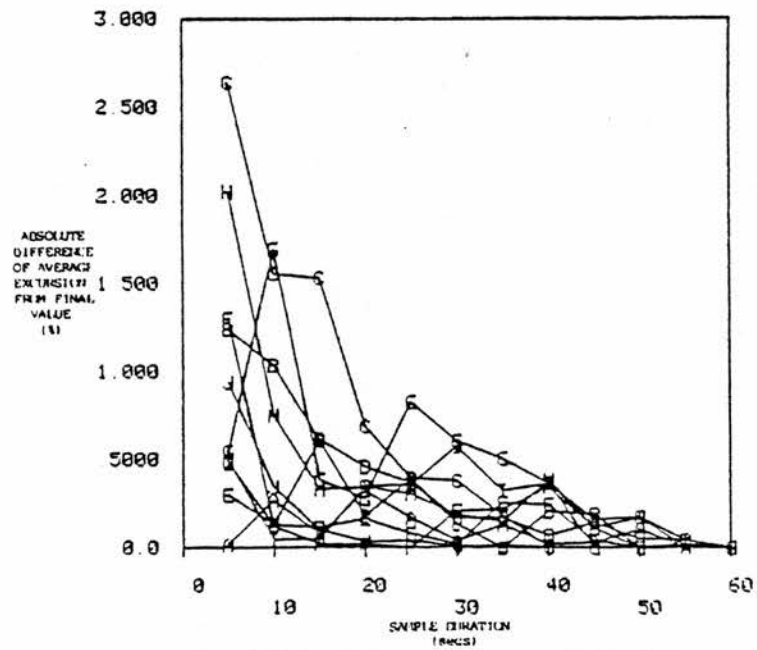


Figure 9 Absolute differences of average excursion from the final long-term value plotted against sample duration (in cumulative 5 second increments) for 10 normal female speakers (labeled A-J).

The threshold criteria are summarized as follows: (1) a difference curve is derived for each speaker on each parameter based on the differences between each cumulative value and the speaker's final long-term value, (2) a threshold is determined for each difference curve based on a percentage of the final long-term value for each curve, (3) a duration for each parameter is chosen based on the requirement that 95% of the speakers pass a given threshold and (4) the choice of threshold is conditioned by the requirement that the difference curve must remain within the given threshold band for at least 10 seconds. Figures 10 and 11 are examples of using the threshold criteria for the mean F0 parameter for the male and female groups, respectively. Each figure is divided into 2 sections where section (a) displays the results for the application of the 1% threshold and section (b) displays the results for the 2% threshold. Each section of the figures is in the form of a bar graph, the abscissa representing sample duration in cumulative 5 second increments and the ordinate representing each individual speaker (labelled A - J as in the previous figures). Each bar depicts when the speaker passed a threshold (i.e. the transition from the unshaded to the shaded region) and the duration over which the difference curve remained within the threshold level (i.e. the transition from the shaded region to the hatched region). The final point in the hatched region marks the overall duration of the speaker's speech sample. The differing thresholds displayed in the figures (i.e. 10a and 11a versus 10b and 11b) represent 2 attempts to determine when 95% of the speakers fell within a given percentage of their final F0 values. For example, male speaker H in Figure 10a displayed a mean F0 difference curve which passed a 1% threshold at 35 secs and remained within that band for 15 seconds. For the 2% threshold shown in Figure 10b, male speaker H passed that threshold at 20 secs and remained within that band for 35 seconds. If all the F0 difference curves in Figures 10a and 11a are examined in this way, one can see that only 70% of the male and female speakers fall within the 1% threshold for the required 10 second duration while all the speakers fulfil the requirements at the 2% threshold (as shown in Figure 10b and 11b). Applying the 95% criterion for group performance to the 2% bars of the male and female speakers, it appears that a minimum 35 sec speech sample would be sufficient to derive long-term mean F0 values to within a 2% accuracy of each speaker's long-term behavior. Therefore, this threshold method suggests the appropriate duration for a given parameter for a given accuracy.

Table 1 summarizes the threshold findings for all the parameters examined in this study. The F0 intonational measures required a 2% threshold to achieve at least 95% agreement amongst the 20 speakers. A duration of 35 secs of oral reading fulfilled all the requirements for the mean F0 for a combined group of 10 male and 10 female speakers. The table also includes a breakdown of intonational results into 2 groups based on gender. The results for the two sub-groups are similar though the female group required 30 secs for stabilization of mean F0 as compared to the 35 sec duration of the male group. Three of the 4 perturbation measures (mean magnitude of excursions, standard deviation of the excursions and DPF) reached the 95% group agreement for a threshold of 10% while the RATEX measure obtained 95% group agreement at the 5% threshold level. The lower threshold level reached by the intonational measures as compared to the perturbation measures reflects a more rapid approach and stabilization of the difference curves towards the final long-term intonational values. Therefore, more speech

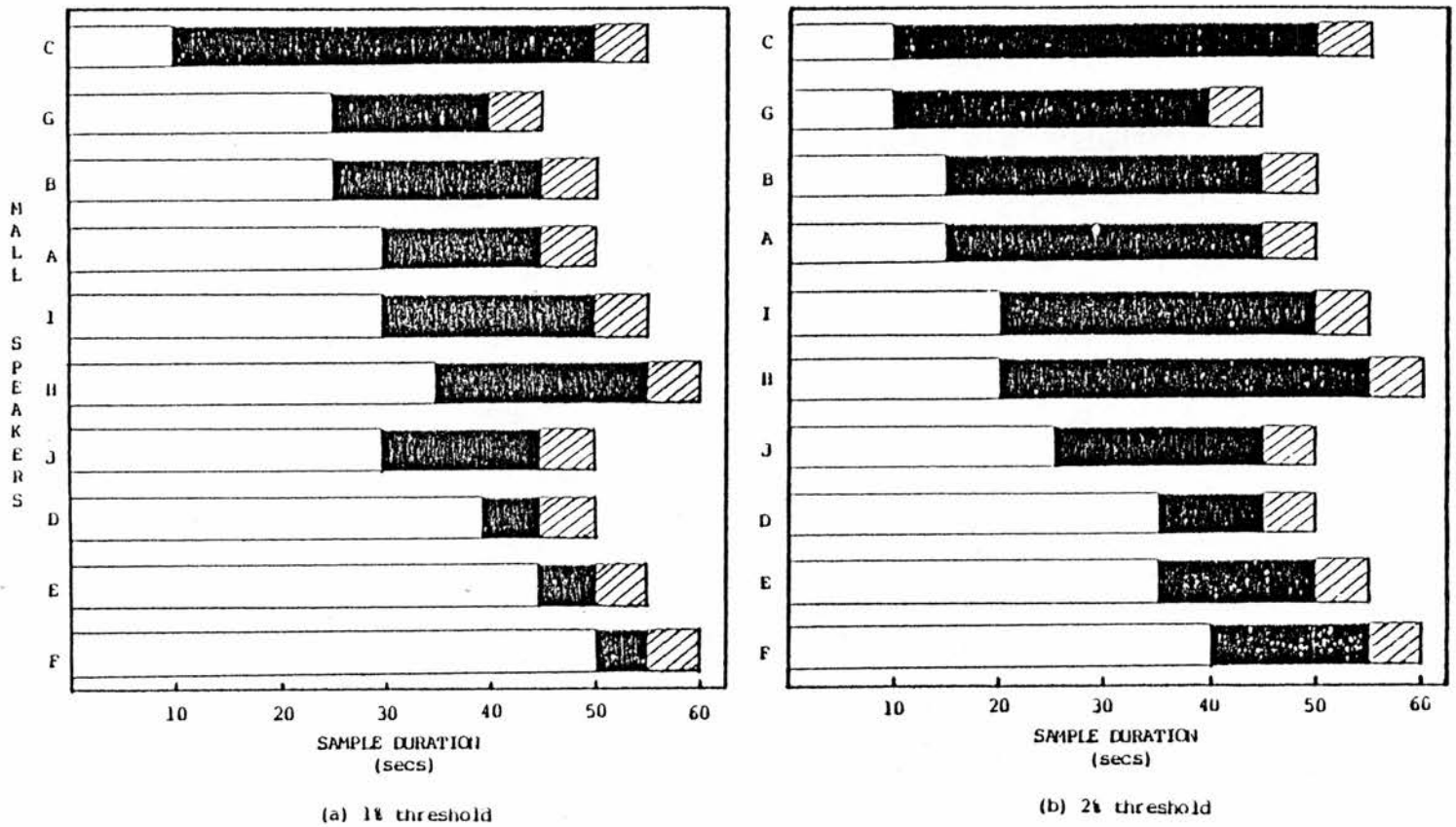


Figure 10 Bar graphs displaying the sample durations at which the male speakers pass 1% or 2% thresholds for mean F0 (Hz) and the periods during which speakers remain below the threshold. Each bar is divided into 3 sections: Unshaded - sample durations before passing the threshold; shaded - sample durations below the threshold; hatched - marks the overall sample duration for each speaker. Figure 9a - 10 normal male speakers (A-J) for the 1% threshold; 9b - 10 normal male speakers for the 2% threshold.

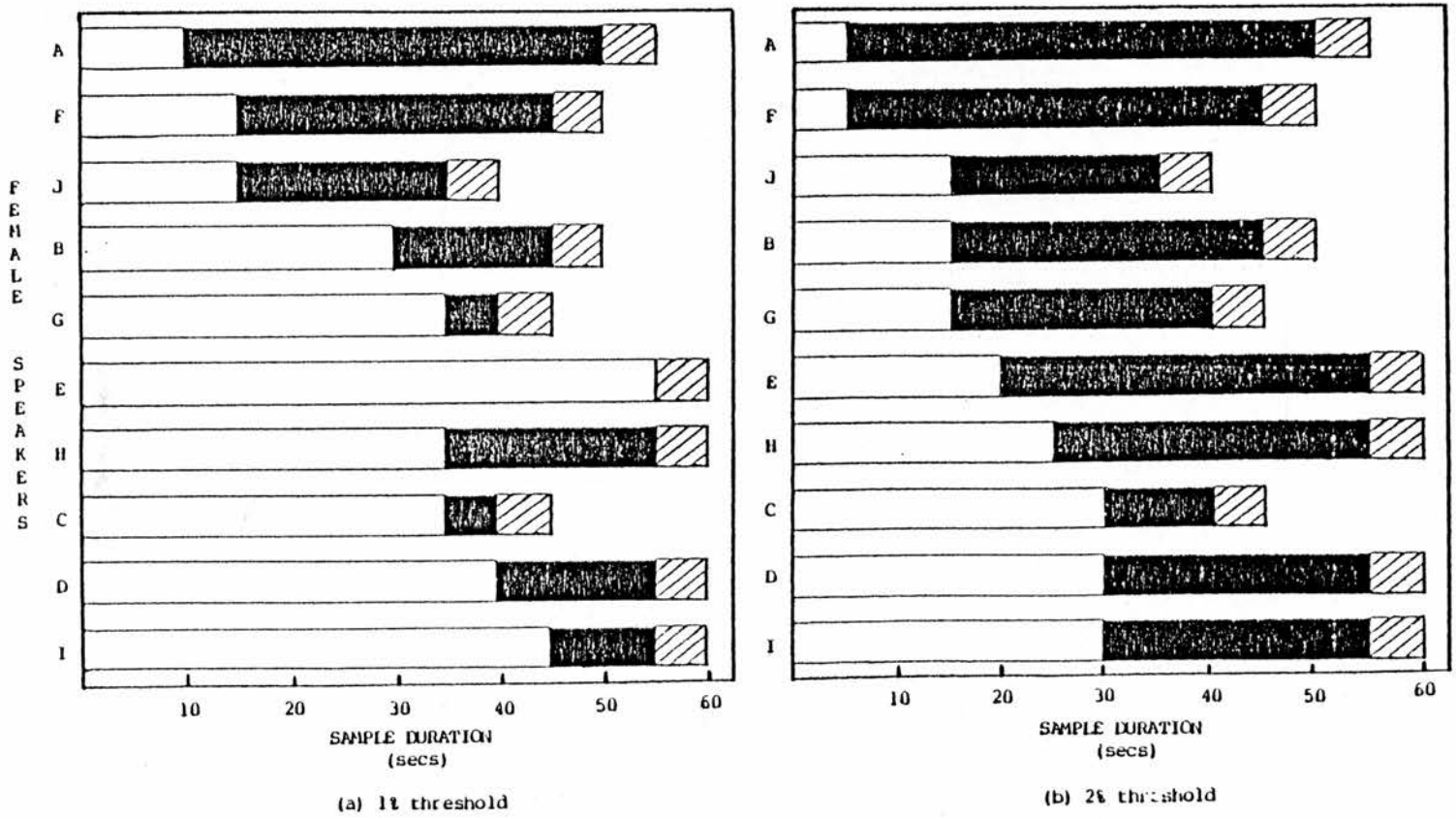


Figure 11 Bar graphs displaying the sample durations at which the female speakers pass 1% or 2% thresholds for mean F0 (Hz) and the periods during which speakers remain below the threshold. Each bar is divided into 3 sections: Unshaded - sample durations before passing the threshold; shaded - sample durations below the threshold; hatched - marks the overall sample duration for each speaker. Figure 10a - 10 normal female speakers (A-J) for the 1% threshold; 10b - 10 normal female speakers for the 2% threshold.

Parameter	N	Threshold (%)	Percent Agreement	Seconds
Mean FO	10M + 10F	1	70	NA
		2	100	35
Mean FO	10M	1	70	NA
		2	100	35
Mean FO	10F	1	70	NA
		2	100	30
St. Dev. FO	10M + 10F	2	70	NA
		5	100	35
St. Dev. FO	10M	2	80	NA
		5	100	35
St. Dev. FO	10F	2	90	NA
		5	100	35
Average Excursion	10M + 10F	5	70	NA
		10	95	35
St. Dev. of Excursions	10M + 10F	5	85	NA
		10	100	30
RATEX	10M + 10F	5	95	40
		10	100	25
DPF	10M + 10F	5	85	NA
		10	100	25

Table 1: This table indicates the sample durations required for each parameter to reach stability based on the application of the threshold and group agreement criteria. NA = not applicable since group did not fulfil 95% group agreement criterion.

data would need to be elicited from a number of the speakers to produce long-term perturbational measures with the accuracy demonstrated for the intonational measures. Having noted this limitation, it can be seen that durations of 25 to 40 secs will produce perturbation measures with substantial accuracy for samples of oral reading from normal speakers. Thus a duration of 40 seconds can be considered a useful practical value for evaluating perturbation and intonation measures derived from samples of oral reading produced by normal speakers. It remains to be established by further research whether this duration would give equally satisfactory results in the evaluation of perturbation and intonation measures derived from speakers with laryngeal pathology. In addition, further research should examine the durational aspects of the perturbation measures for a variety of speaking tasks including differing speech samples and replicability of a given task.

Conclusions

Forty second samples of read speech should provide relatively stable long-term speaker-characterizing parameters of intonation and perturbation in normal speakers. This finding is in general agreement with the results of previous studies of the long-term features of the voice. Comparable durations of speech samples are required to produce stable long-term voice parameters from normal female and male speakers. For the highest levels of accuracy, perturbation measures would require longer durations of speech than do intonational measures. Further research is required to determine these durations for the more accurate perturbation measures. The development of durational growth curves for the intonational and perturbational measures may provide a useful indicator of voice function.

References

Askenfelt, A. and Hammarberg, B. (1981). 'Speech waveform perturbation analysis revisited'. Stockholm: Speech Transmission Lab., Quarterly Progress and Status Report, 4, 49-68.

Atal, B. (1976). 'Automatic recognition of speakers from their voices'. Proceedings, IEEE, 64, 460-475.

Davis, S. B. (1979). 'Acoustic characteristics of normal and pathological voices'. In Lass, N. J. (ed.), Speech and Language: Advances in Basic Research and Practice, New York: Academic Press, 1, 273-338.

Fairbanks, G. (1960). Voice and Articulation Drillbook. New York: Harper Brothers.

Gold, B. and Rabiner, L. (1969). 'Parallel processing techniques for estimating pitch periods of speech in the time domain'. JASA, 46, 442-448.

Green, N. (1972). 'Automatic speaker recognition using pitch measurements in conversational speech. JSRU Report No. 1000', Joint Speech Research Unit, Ruislip, Middlesex.

Hanson, R. J. (1978). 'A two-state model of FO control'. JASA, 64, 543-544.

Hecker, M. and Kreul, E. (1971). 'Descriptions of the speech of patients, with cancer of the vocal folds. Part 1 : measures of fundamental frequency'. JASA, 49, 1275-1282.

Hess, W. (1983). Pitch Determination of Speech Signals - Algorithms and Devices. Berlin: Springer-Verlag.

Hiller, S. M., Laver, J. and Mackenzie, J. (1983). 'Automatic analysis of waveform perturbations in connected speech'. Edinburgh University Department of Linguistics, Work in Progress, 16, 40-69.

Holmes, J. N. (1973). 'The influence of glottal waveform on the naturalness of speech from parallel formant synthesizer'. IEEE Trans. Aud. and Electroac., AU-21, 298-305.

Horii, Y. (1975). 'Some statistical characteristics of voice fundamental frequency'. JSHR, 18, 192-201.

---------- (1979). 'Fundamental frequency perturbation observed in sustained phonation'. JSHR, 22, 5-19.

Laver, J., Hiller, S. and Hanson, R. (1982). 'Comparative performance of pitch detection algorithms on dysphonic voices'. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Paris 1982, 192-195.

Laver, J., Wirz, S., Mackenzie, J., and Hiller, S. (1981). 'A perceptual protocol for the analysis of vocal profiles'. Edinburgh University Department of Linguistics, Work in Progress, 14, 139-155.

Lieberman, P. (1963). 'Some acoustic measures of the fundamental periodicity of normal and pathological larynges'. JASA, 35, 344-353.

Markel, J. D. and Davis, S. B. (1979). 'Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base'. IEEE Trans. Ac., Sp. and Sig. Proc., ASSP-27, 74-82.

Markel, J., Oshika, B. and Gray, G. (1977). 'Long-term feature averaging for speaker recognition'. IEEE Trans. Ac., Sp. and Sig. Proc., ASSP-25, 330-337.

Mead, K. O. (1974). 'Identification of speakers from fundamental frequency contours in conversational speech'. JSRU Report No. 1002, Joint Speech Research Unit, Ruislip, Middlesex.

Nolan, N. (1983). The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press.

Rabiner, L., Sambur, M. R. and Schmidt, C. E. (1975). 'Applications of nonlinear smoothing algorithm to speech processing'. IEEE Trans. Ac., Sp. and Sig. Proc., ASSP-23, 552-557.

APPENDIX 6

Proceedings of The Institute of Acoustics

ACOUSTIC ANALYSIS OF VOCAL FOLD PATHOLOGY

John Laver, Steven Hiller, Janet Mackenzie

Centre for Speech Technology Research
Department of Linguistics, University of Edinburgh.

INTRODUCTION

The main aim of this study is to develop a computer-based system of acoustic analysis which is capable of screening voices for the presence of vocal pathologies. The social implications of such a system, which involves a non-invasive and relatively cheap recording technique, are considerable. The early detection of such disorders as laryngeal cancer is highly desirable, since prompt medical treatment has a high success rate.

Acoustic screening can be profitably applied to two different populations. The first of these is an unselected population. For example, routine screening for vocal pathology could be carried out in "well woman" or "well man" clinics, alongside existing screening tests for breast cancer, cardiac function etc. Alternatively, a more limited, pre-selected population could be screened in cases where vocal pathology is already suspected. For instance, patients referred by their general practitioners for laryngeal examinations may face lengthy waiting lists. If these patients were to be recorded at the time of referral, it might be possible to ensure that those cases where acoustic measures suggest the presence of a neoplastic structural abnormality would be seen immediately. Acoustic screening could thus be used to select priority cases, accelerating examination of patients judged to be at serious risk.

A second aim of the project is to investigate the possibility of using acoustic measures to differentiate between various types of vocal pathology. This requires an exploration of the relationships between various classes of structural disturbance of the vocal fold tissue layers and acoustic perturbations of the laryngeal waveform. It is possible to formulate a range of predictions about the acoustic consequences of different types of pathology [1], and these predictions are being tested by collecting detailed information about the status of the individual patient's larynx. This is made possible by collaboration with laryngologists and speech therapists at the Radcliffe Infirmary, Oxford and the Royal Infirmary, Edinburgh.

THE ACOUSTIC SYSTEM

The research results to be discussed derive from computer-based acoustic measurement of individual pitch periods in approximately 40 seconds of tape recorded read text. The measurement system uses an elaborated version of the Gold and Rabiner parallel processing method [2], with phase compensation, low-pass filtering, non-linear smoothing for intonational baseline measurement, and sampling frequency multiplication at waveform peaks by parabolic interpolation for greater resolution of individual pitch periods. This system is described in detail in Hiller, Laver and Mackenzie [3]. There are three types of output data. Firstly, intonational data, such as mean, median and standard deviation of fundamental frequency (FØ), is obtained from the smoothed FØ trend line. Secondly, statistical analysis of excursions of individual pitch period values from the local fundamental frequency trend line provides data for pitch perturbation. Thirdly, an analysis of intensity perturbation is made. The following

(Proceedings of the Institute of Acoustics, 6, 235-452, 1984)

measures are used to describe both pitch and intensity perturbation:

1. Mean magnitude of excursion
2. Standard deviation of the magnitude of excursion
3. The rate of excursion (RATEX). This is the percentage of points in the sample where the magnitude of excursion is equal to, or more than 3% of the local trend line value.
4. The directional perturbation factor (DPF). This measure, adapted from Hecker and Kreul [4] is the percentage of changes in algebraic sign in the excursion values. A 3% threshold is also applied to this measure.

SUBJECTS, SPEECH MATERIAL AND ANALYSIS PROCEDURES

Table 1 presents information about sex, age and the percentage of self-reported smokers in the control group and pathological group investigated in this study. It can be seen that the control group speakers are on average younger than the pathological speakers. This reflects a bias in collection of control speakers, who were mostly selected from the university environment. Future work in this project will rectify this bias. None of the control speakers reported any known speech or hearing problems (including cold or sinus ailments) at the time of recording.

Table 1. Subject group information

Group	Sex	N	Mean age (range)	Percentage smokers
Control	M	38	34 (18-63)	19.4%
Control	F	26	26 (18-44)	29.2%
Pathological	M	32	54 (27-82)	30.3%
Pathological	F	31	56 (26-75)	51.7%

Table 2 shows a broad classification of the laryngeal disorders evidenced by the pathological speakers, as determined by laryngological examination.

Table 2. Classification of Laryngeal disorders diagnosed in pathological group

Type of pathology	Number of males	Number of females
Epithelial disorders (e.g. carcinoma, papilloma keratosis)	12	0
Polyps, nodules	5	10
Disorders of the cartilagenous area	4	3
Mild oedema, redness etc	5	14
Palsies	6	4
Total	32	31

A tape recording was made of each speaker as he or she read the first two paragraphs of "The Rainbow Passage" [5]. Forty seconds of each recorded speech sample was digitized at 20 KHz and stored on computer magnetic tape for future processing by the acoustic system. The results of the analysis were stored in a computerized voice acoustic data base for use in screening procedures.

SCREENING APPROACHES AND RESULTS

There are several possible approaches to acoustic screening for vocal pathology. One simple approach is to focus on single acoustic parameters, and to relate individual values to the means and standard deviations (SD) of control group data. This is easiest if parameter values are normalised by conversion to Z-scores. In other words, they are expressed in terms of units of SD by which they diverge from the control group mean. It seems reasonable to suppose that any individual parameter value which diverges from the control group mean by more than 2SD is indicative of a strong risk of abnormality. The cut-off level for screening can therefore be set at 2SD above or below the control group mean for a given acoustic parameter.

A preliminary evaluation of the screening potential of 10 individual acoustic parameters, using this approach, shows a considerable range of success. This is shown in Table 3. The most effective parameter, in terms of discrimination between control speakers and those with known laryngeal pathology, is shimmer DPF. This picks out 62.1% of male speakers and 53.6% of female speakers with known pathology. The least effective parameter is SD of shimmer excursions, for which only 6.9% of pathological males and 7.1% of pathological females have values which diverge from the control mean by more than 2SD.

Table 3. Percentage of speakers with acoustic values which diverge from the control group mean by 2SD

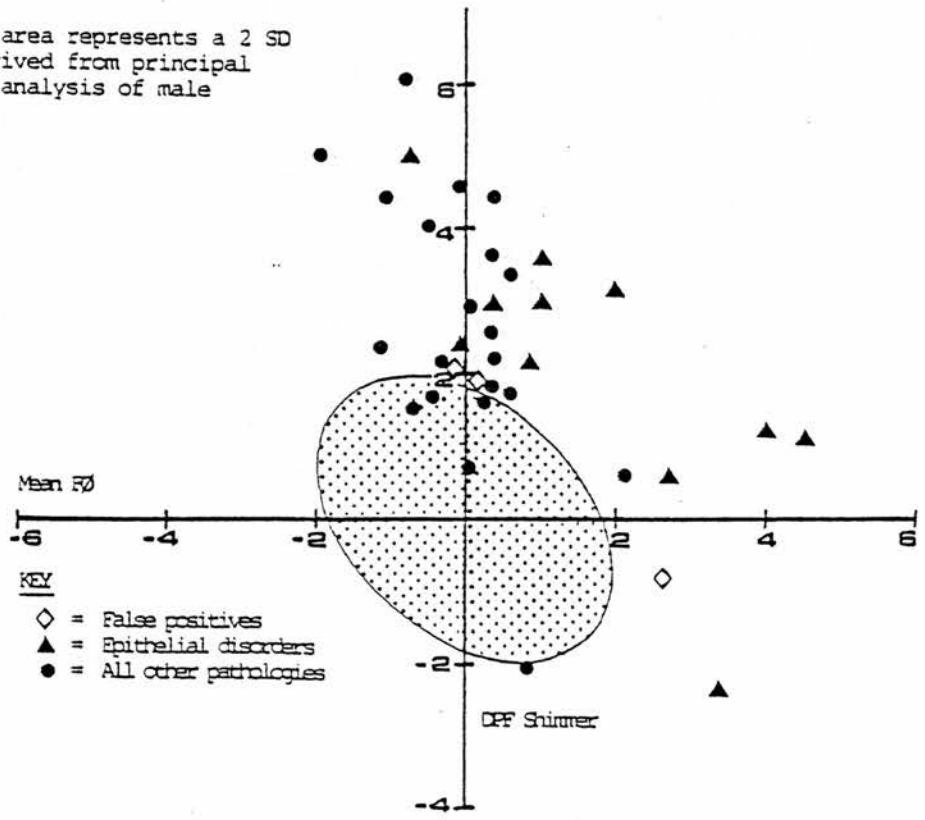
Acoustic parameter	Male controls	Male pathological	Female controls	Female pathological
Shimmer DPF	... 2.6%	62.1%	3.8%	53.6%
Shimmer SD of excursions	... 5.3%	6.9%	7.7%	7.1%
One or more out of ten parameters ..	.. 18.4%	82.8%	30.8%	92.9%
Two or more out of ten parameters .	.... 2.6%	69.0%	11.5%	78.6%

An alternative criterion for screening for pathology might be divergence from the control group mean by more than 2SD in at least one of the overall set of 10 acoustic parameters. This gives a rather better detection rate for speakers with laryngeal pathology (82.8% of males and 92.9% of females), but it also identifies an unacceptably high number of the control group as being apparently abnormal (18.4% of males and 30.8% of females). We can call these "false positives". A compromise criterion for abnormality is that at least 2 parameters out of 10 must diverge from the control group means by more than 2SD. This reduces the false positives to 2.6% of males and 11.5% of females, whilst still detecting 69.0% of males and 78.6% of females with known laryngeal pathology.

Although mean FØ alone does not seem to be a very good screening parameter, it is interesting in that it was the only parameter which was abnormal in one early case of laryngeal cancer. This prompts the hypothesis that in some cases patients may be able to maintain normal levels of perturbation in spite of slight structural abnormality, but only within a rather abnormal pitch range. It therefore seems useful to examine the data in terms of the conjunction between mean FØ and perturbation, to see if this forms a better basis for screening. (A plot of mean FØ against perturbation has the added advantage that it can be related to predictions about the probable consequences for FØ of alterations in stiffness, mass and symmetry of the vocal folds). We can concentrate initially on shimmer DPF, as the perturbation measure which was most effective as a single screening parameter.

On the graph shown in Figure 1, the intersection of the axes at zero corresponds to the control group mean for each parameter. Each unit away from the mean corresponds to one standard deviation. By definition, the control group will cluster around the intersect.

Figure 1: A scattergram of DPF shimmer vs. Mean FØ for male speakers



A principal components analysis can be applied to the control group data, and ellipses can be drawn which indicate the covariance between the two parameters. If an ellipse is drawn at the 2SD level, this will, by definition, enclose approximately 95% of the control group. This ellipse can then be used as a screening boundary, so that any data point falling outside the ellipse is interpreted as being potentially indicative of abnormality.

The 2SD ellipse shown on Figure 1 describes the male control group data. Of the 38 control males, only 3(7.9%) fall outside the ellipse, and would thus be picked up as false positives. In contrast, if the male speakers with laryngeal pathology are plotted on this graph, 87.5% fall outside the ellipse. This type of two dimensional analysis thus looks to be a more promising approach to screening than a single parameter approach.

It should be remembered that this 87.5% detection rate applies to a group which includes a wide range of disorders, including some very minor and benign pathologies. Laryngeal cancer and pre-cancerous states, which are obviously the most serious disorders, almost always arise in the covering epithelium, so that we are most concerned with detection of structural tissue alterations of the epithelium. It is encouraging that 100% of this subgroup fall outside the ellipse.

The results for female speakers show a similar pattern, with 15.4% of controls, and 90.3% of pathological speakers falling outside a 2SD ellipse. The slightly higher percentage of false positives may reflect the smaller female control group, which does not yet show a normal distribution of mean FØ values.

It is worth commenting that shimmer DPF does not, however, seem to be the best acoustic measure for differentiation between classes of vocal pathology.

There is some evidence that a plot of jitter Ratex against mean FØ, whilst having less success in initial screening, is a better discriminator for pathology type. For example, there is a clear tendency for epithelial disorders to result in high jitter ratex scores and/or high mean FØ, whereas a large proportion of nodules and polyps, have low jitter ratex scores and low mean FØ compared with the control group.

A more sensitive, but statistically more complex, approach to acoustic discrimination of different vocal pathologies would involve a comparison of overall acoustic profiles. An acoustic profile would include information on all the acoustic measures produced by this system. The average profile shapes for different classes of vocal pathology are recognisably different, and this supports a further investigation of this approach.

CONCLUSION

In conclusion, these results appear to justify attempts to use acoustic analysis in screening for the presence of laryngeal pathology. The task of acoustic discrimination between types of pathological abnormality is more complex, but there are indications that acoustic analysis may provide useful quantitative information in support of the diagnostic process.

REFERENCES.

[1] J. Mackenzie, J. Laver and S. Hiller, 'Structural pathologies of the vocal folds and phonation'. Work in Progress, Dept.of Linguistics, Edinburgh University, No. 16, 8-116, (1983).

[2] B. Gold and L.R. Rabiner, 'Parallel processing techniques for estimating pitch periods of speech in the time domain', J.A.S.A., Vol. 46, 442-448, (1969).

[3] S. Hiller, J. Laver and J. Mackenzie, 'Automatic analysis of waveform perturbations in connected speech'. Work in Progress, Dept. of Linguistics, Edinburgh University, No. 16, 40-68, (1983).

[4] M. Hecker and E. Kreul, 'Descriptions of the speech of patients with cancer of the vocal folds. Part 1: Measures of fundamental frequency'. J.A.S.A., Vol. 49, 1275-1282, (1971).

[5] G. Fairbanks, Voice and Articulation Drillbook, New York: Harper Brothers, (1960).